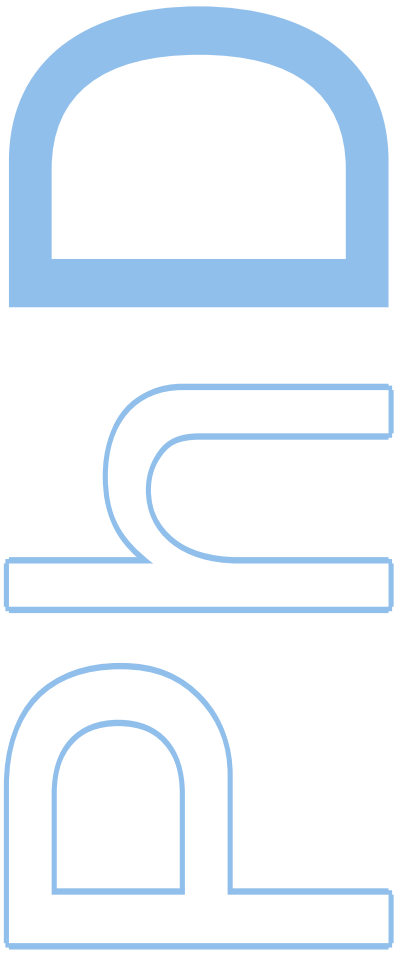
Modelling Reporting Delays for Infectious Diseases: An Application to the Portuguese HIV - AIDS Surveillance Data



Alexandra Alves Oliveira

PhD Programme in Applied Mathematics

Department of Mathematics

2018

Supervisor

Professora Doutora Ana Rita Gaio, Assistant Professor,

Department of Mathematics, Faculty of Sciences of the University of Porto

Co - Supervisor

Professor Doutor Luís Paulo Reis, Associate Professor,

Department of Informatics Engineering, Faculty of Engineering of the University of Porto

Modelling Reporting Delays for Infectious Diseases: An Application to the Portuguese HIV-AIDS Surveillance Data

by

Alexandra Alves Oliveira

A thesis submitted in conformity with the requirements
for the degree of Doctor of Applied Mathematics
Department of Mathematics, Faculty of Sciences, University of Porto

Dedication

Em memória do meu avô Álvaro, que me ensinou o poder regenerador e protetor da Alegria, e da minha avó Margarida que me ensinou o poder transformador do Trabalho.

Para os meus pais Tina e Zé por me ensinarem que o Amor e União movem montanhas e para o meu marido Carlos por o viver e partilhar comigo todos os dias!

Para ti filhotinho, por todos os teus raios de sol que me completam e são a cor do meu coracão.

Acknowledgements

This was a long journey that encompassed much more than technical knowledge alone. As it should, being a doctoral candidate, I had to develop a profound understanding of the reality, along with its limitations and dynamics, connect with other researchers around the world, explore other knowledge fields and perspectives, and many other challenges that made me evolve and learn to trust myself and my work, while searching for answers and solutions.

This work would never be finished if it wasn't for the technical supervision, support and inspiration of Professor Rita Gaio and Professor Luís Paulo Reis for mentoring and encouragement me towards the right strategy and enforcing the opportunities in the midst of the unknown.

Therefore I am thankful to Professor Henrique Barros, former Coordinator of Portuguese Commission of HIV - AIDS infection, for sparking the interest on the problem and for all the useful insights based on his precious field experience. European Centre of Disease Control, namely through Anastasia Pharris, Chantal Quinten and Joana Gome Dias from the HIV - AIDS surveillance network by caring about my research, enabling and encouraging me to work with larger datasets from European countries, while understanding their contexts and Public Health policies. Also, Magdalena Rosińska of the The Narodowy Instytut Zdrowia Publicznego – Państwowy Zakład Higieny by all the support, encouragement and useful discussions of the European context in terms of reporting delay.

I would also like to express my gratitude to Mark Van Der Loo, from Statistics Netherlands, whose important work towards Statistical Data Cleaning shed a light on my research while struggling with the real world datasets, due to all the information noise that still prevails. Also, Peter Fader from the Wharton School of the University of Pennsylvania and Bruce Hardie from the London Business School, for transforming complexity into meaningful knowledge and words of advice about form modelling underreporting in Poisson processes.

Finally, to my family and friends that so many times defended the silver lining and help me go through and beyond my transformation and fears I dedicate part of a musical poem that me and my husband shared when we met:

Pain throws your heart to the ground
Love turns the whole thing around
Fear is a friend who's misunderstood
But I know the heart of life is good
I know it's good
John Mayer

Abstract

In a deeply interconnected world of people and goods, infectious diseases constitute a serious threat. For a prompt and effective response, it is crucial to have accurate and timely information. These two characteristics are related with two main lines of research: data quality assessment and improvement and modelling of the reporting delay process.

Every case-reporting system has some degree of under-ascertainment. Timeliness is a widely acknowledged key indicator for surveillance data quality, in particular the time between diagnosis and report to the national surveillance system – the reporting delay. A particular case of a surveillance system is the HIV–AIDS surveillance system. It differs from others due to the special transmission modes of the disease, charged with heavy social stigma, long asymptomatic or with mild symptoms latency period of the infection, lack of affordable treatment and cure, and high fatality rates.

In this thesis we explore many of the breakthrough new ideas in statistical analysis for modelling the complex reality of reporting delays in a surveillance system, emphasizing the statistical methods and their conceptual underpinnings.

It is presented a holistic description of epidemics by infectious diseases and how they are surveyed and reported. The quality of the data is assessed by a new reproducible methodology . The methodology is based on statistical data pre-processing of surveillance systems data, and the data quality definition used is an extension of the ISO / IEC 25012, including statistical dimensions.

Also, the application of a statistical model for estimation of the under-reporting phenomenon in the Portuguese epidemic is assessed as well as the changes on the reporting delay over time, identifying the main factors influencing the reporting delay and considering the individual, administrative, organizational and historical contexts.

The classical joint modelling with Poisson counts does not fully capture the reporting process of the Portuguese HIV -AIDS system. These results are consistent with a previous publication using data from the same system. Fluctuations in reporting delay patterns over time were found which can be attributed to changes in the distributions of the covariates, or to temporary periods of slower reporting in specific geographical areas.

The reporting delay was divided into quarters and in a 2-class division with a cut-off at the 3-months delay and several supervised learning techniques were applied in order to identify the main factors influencing these classes. Also, reporting delays were modelled in continuous time and with a fully multilevel parametric approach, allowing for inferences on covariate effects to be mainly driven by the shorter delays, which are appropriate when the goal is to rapidly detect changes in epidemic patterns. This model found a positive effect of the risk group and of the reporting entity work load and a negative effect of stage and age.

We hope the present study may be useful for the definition and implementation of future surveillance systems, highlighting the need to empower the data quality alongside with new statistical tools and techniques to better survey our diseases.

Keywords: Infectious Diseases, HIV-AIDS, Surveillance System, Surveillance Data, Statistical Pre-Processing, Multilevel Modelling, GLMs, Data Mining, Reporting Delay, Under-Reporting

Resumo

Num mundo profundamente interligado de pessoas e bens, as doenças infeciosas constituem uma séria ameaça. Para garantir uma resposta rápida e eficaz, é crucial ter informações precisas e oportunas. Essas duas características podem estar na base de duas linhas principais de investigação: a avaliação de qualidade do dados, assim como a melhoria dos mesmos, e análise dos processos de atraso na notificação de casos.

Cada sistema de vigilância tem algum grau de sub-averiguação sendo a prontidão um indicador chave amplamente reconhecido para a qualidade dos dados, em particular o atraso na notificação de um caso diagnosticado para o sistema de vigilância nacional - o atraso na notificação. Um caso particular de um sistema de vigilância nacional é o VIH - SIDA. Este difere de outros sistemas de vigilância devido a fatores como o modo de transmissão da doença, carregados de forte estigma social, longo período de latência assintomático ou com sintomas leves, a falta de tratamento e de cura acessíveis e as altas taxas de mortalidade.

Nesta tese exploram-se várias metodologias recentes de análise estatística de atrasos de notificação num sistema de vigilância de uma doença infeciosa. Neste enquadramento são enfatizados os métodos estatísticos úteis no contexto particular do VIH - SIDA e os seus fundamentos conceituais. É apresentada uma descrição holística das epidemias por doenças infecciosas e como elas são vigiadas e notificadas. É também proposta e implementada uma metodologia reprodutível para a avaliação e melhoria da qualidade dos dados estatísticos provenientes de sistemas de vigilância, usando como definição de qualidade de dados uma extensão da ISO / IEC 25012, que inclui dimensões estatísticas. É efetuada a avaliação da aplicação de modelos estatísticos para estimação do fenómeno de subnotificação da epidemia em Portugal e as alterações no atraso na notificação ao longo do tempo. Finalmente são identificados os principais fatores que influenciam o atraso do relatório, tendo em consideração o contexto individual, administrativo, organizacional e histórico.

A modelação clássica baseada em contagens com distribuição de Poisson não descreve completamente o processo de notificação no sistema Português de vigilância do VIH - SIDA. Estes resultados são consistentes com publicações anteriores para o mesmo sistema. Foram encontradas flutuações ao longo do tempo nos padrões de atraso na notificação que podem ser atribuídos a mudanças nas distribuições das covariáveis ou a períodos temporários de notificação mais lentos e em áreas geográficas específicas. O atraso na notificação foi dividido em trimestres e foi considerada uma divisão em 2 classes com um ponto de corte de 3 meses. Várias técnicas de *data mining* foram aplicadas para identificar os principais fatores que influenciam estas classes. Além disso, os atrasos na notificação foram modelados em tempo contínuo e com uma abordagem paramétrica hierárquica, permitindo que as inferências sobre efeitos de covariáveis sejam impulsionadas principalmente pelos atrasos mais curtos, que são apropriados quando o objetivo é detetar rapidamente mudanças nos padrões epidémicos. Este modelo identificou um efeito positivo do grupo de risco e da carga de trabalho da entidade notificadora e um efeito negativo do estado da doença e da idade.

Esperamos que o presente estudo possa ser útil para a definição e implementação de futuros sistemas de vigilância, destacando a necessidade de avaliar e melhorar a qualidade dos dados juntamente com novas ferramentas e técnicas estatísticas.

Palavras-Chave: Doenças infeciosas, HIV-SIDA, Sistema de Vigilância, Dados de Vigilância, Pré-processamento Estatístico, Modelagem multinível, GLMs, Mineração de Dados, Atraso na Notificação, Subnotificação

Contents

List of Tables

List of Figures

Abreviations

ACSS Central Administration of the Health System.

AFT Accelerated Failure Time Models.

AIC Akaike Information Criterion.

AIDS Acquired Immunodeficiency Syndrome.

ART Anti-Retroviral Therapy.

BCG Bacillus Calmette-Guérin.

BIC Bayesian Information Criterion.

CART Classification and Regression Trees.

CDC Centers for Disease Control and Prevention.

CDisF Cumulative Distribution Function.

CNLCS Comissão Nacional de Luta Contra a SIDA.

CNSida Coordenação Nacional para a Infecção VIH - SIDA.

CVEDT Centro de Vigilância Epidemiológica das Doenças Transmissíveis.

ECDC European Centre of Disease Control and Prevention.

EEA European Economic Area.

EM Expectation Maximization.

ESS QAF The Quality Assurance Framework of the European Statistical System.

EU European Union.

FSWs Female Sex Workers.

GLMs Generalized Linear Models.

HA Hungarian Algorithm.

HAART Highly Active Anti-Retroviral Therapy.

HIV Human Immunodeficiency Virus.

IDU Injecting Drug User.

IDUs Injecting Drug Users.

INE Instituto Nacional de Estatística.

KS Kaposi's Sarcoma.

LVT Lisbon and Tagus Valley.

MCMC Markov Chain Monte Carlo.

MDR-TB Multidrug-Resistant Tuberculosis.

MSMs Men who have Sex with Men.

MTB Mycobacterium tuberculosis.

r.t.h. Reverse Time Hazard Function.

RHA Regional Health Administration.

SA Sampling Algorithm.

SC Supervised Clustering.

SEM Socio-Ecological Model.

SSPRM Semi-Supervised Probabilistic Relational Models.

SVM Support Vector Machines.

SWs Sex Workers.

TB Tuberculosis.

TESSy The European Surveillance System.

UNAIDS Joint United Nations Programme on HIV/AIDS.

WHO World Health Organization.

WHOE World Health Organization - Regional Office for Europe.

Chapter 1

Introduction

No man is an island,
Entire of itself,
Every man is a piece of the continent,
A part of the main. John Donne

1.1 Context

In a deeply interconnected world of people and goods, infectious diseases constitute a serious threat. So, an active vigilance "for signs of an outbreak, rapid recognition of its presence, diagnosis of its microbial cause" is required [1]. It is also necessary to identify modes of transmission and at-risk population groups, and to define public measures for target prevention [2, 3]. The collection of adequate data is vital to evaluate the burden of the disease and the impact of prevention and control programmes, aiming at effective and efficient responses.

Typically, this implies a complex system that relies on processes and individuals and thereby can be found to differ substantially according to the disease, health condition and country. Also, traditional public health surveillance approaches use pre-specified case definitions and employ manual data collection, human decision making, and manual data entry [4, 3, 5, 6].

Every case-reporting system has some degree of under-ascertainment. This can be due to failure of patients to present for diagnosis, failure of physicians to diagnose or report the disease, and failures in the health department itself to count cases owing to misclassification or other reasons [7]. Timeliness is another widely - acknowledged key indicator for surveillance data quality. It should be periodically evaluated because it can reflect the reporting time delay between any number of response steps in a surveillance process [3].

Reporting delays may depend on a number of factors such as the patient's recognition of symptoms; the patient's acquisition of medical care; the use of confirmatory laboratory testing; reporting by the health care provider or the laboratory to the local, region, or state public health authority; the volume of cases identified in the state; case follow-up investigations to verify the case report or to collect additional case information; periods of decreased surveillance system activity due to variable employment levels; computer system down-time for maintenance, upgrades, or new application development; and data processing routines, such as data validation or error checking [5, 8].

A particular case of a surveillance system is that for the Human Immunodeficiency Virus (HIV) - Acquired Immunodeficiency Syndrome (AIDS). It differs from other surveillance systems in many ways, namely by reflecting the special transmission patterns charged with heavy social stigma, the long asymptomatic or with mild symptoms latency period of the

infection, the lack of affordable treatment and cure, and its high fatality rates [9].

1.2 Motivation

Infectious diseases have been an ever-present threat to mankind and plait the history. From the Biblical Plagues and the Plague of Athens in ancient times, to the Black Death of the Middle Ages, the 1918 “Spanish Flu” pandemic, Tuberculosis, and more recently, the HIV - AIDS pandemic, Cholera, Dengue and Zika, infectious diseases have continued to emerge and re-emerge in a manner that defies accurate predictions [10, 11].

Infectious diseases are the second leading cause of death and disability - adjusted life years worldwide (1 disability - adjusted life year is 1 lost year of healthy life). Over time many microbes, such as `Mycobacterium tuberculosis`, have continued to develop resistance to drugs defying conventional therapies and posing a threat to public health [11]. Another example arises from the epidemiological history of the HIV - AIDS that led to a clear understanding that reducing its incidence is a complex battle whose scope goes way beyond the shaping of the individuals disease related risk-behaviour. A holistic intervention, considering all the main factors, is essential.

Humans are hard to model and in the context of this thesis several layers of complexity are added and need to be addressed. The HIV - AIDS is a complex disease in terms of biology and social context. It is characterized by long and variable asymptomatic or with middle symptoms phases and transmission modes grounded in highly controversial behaviours. These behaviours may cause discrimination and stigma leading the individuals to avoid testing, treatment and / or provide all the necessary information for treatment and surveillance. On the other hand, patients are emerged in large and complex health and surveillance systems composed by bureaucratic processes and, of course individuals. Among these individuals, the health care providers play a crucial role and their main focus is on saving lives not bureaucratic paper work. So, when public health systems report the numbers of new infected individuals they are reporting the disease history but also reflecting the processes of detecting and registering of the cases.

Focusing only on the reporting systems, it is widely recognized that these systems are accommodated by several problems that imprint on data several data quality issues. Most common sampling and non-sampling errors are caused by problems such as [12, 13, 3, 14, 15, 16]:

- unrepresentativeness or selection bias of the patients;
- system implementation, that suffered alterations throughout times, due to governmental rules and regulations;
- case definitions used inconsistently by the main diagnosis stakeholders;
- data processing errors;
- reporting delays.

For a rapid reaction to an emerging infectious disease, it is crucial to have accurate and timely information. These two characteristics can create two main lines of research: data quality assessment and improvement of the raw data, and modelling of the reporting delays process.

Statistical modelling in science remains, partly at least, an art [17]. In practice, there are some principals to guide the analyst and a multidisciplinary approach is often useful. Although a simplification of the reality, statistical models try to mimic the structure of the data and guide the investigation lighting the main relationships among variables.

The field of Statistics is constantly challenged by the problems brought by the real world. With the advent of technology, statistical problems have exploded both in size and complexity [18]. Challenges arise in terms of storage, organization, quality of information and even in relevance of the interpretation of traditional p-values.

A large amount of data is being generated in many fields, and the statistician's job is to make sense of it all: extract relevant information, extract important patterns and trends, and understand "what the data says" [18].

These challenges led to an explosion of methods in the statistical sciences along with new technologies, information systems, decision support systems and clinical parameters prediction algorithm, in particular, in many Health related areas [19].

There are essentially two goals in statistical analysis: prediction of an outcome measure based on a number of input measures and/or description of the associations and patterns among a set of input measures.

This thesis is an attempt to bring many of the breakthrough new ideas in statistical analysis for modelling the complex reality of reporting delays in a surveillance system, emphasizing the statistical methods and their conceptual underpinnings [18].

1.3 Objectives

The aim of the thesis is to study the phenomenon of reporting delay of a surveillance data using all possible information. The main hypothesis addressed in this thesis is:

It is possible to describe the phenomenon of reporting delay, in a surveillance system, modelling the individual factors, as well as the health care and surveillance systems.

In order to accomplish this goal, the following specific objectives were defined:

- to review the relevant related work in the field of modelling reporting delay distributions;
- to survey of the relevant statistical models used within the framework of this thesis;
- to propose and implement a reproducible methodology for assessment and improvement of data quality through statistical data pre-processing for surveillance systems data;
- to evaluate the application of traditional methodologies developed for large epidemics in the Portuguese epidemic, which is small and concentrated;
- to assess the application of a statistical model for estimation of the under-reporting phenomenon in the Portuguese epidemic;
- to assess the changes on the reporting delay throughout time (study the possibility of non-stationarity of the reporting delay process) taking into consideration the nested individual, health care and organizational contexts.;
- to identify the main factors influencing reporting delays of HIV - AIDS cases within the Portuguese surveillance system;
- to compare the models' achieved results and evaluate their performance.

1.4 Contributions

This thesis makes the following contributions to the field:

- It presents a holistic description of HIV -AIDS epidemic and how they are surveyed and reported.
- Unification of the definitions of data quality based on the classical ISO / IEC 25012 and the quality assurance framework of the European Statistical System, to be used by others in related works. Also, it updated the map between the main data set issues affecting each data quality dimension. This work is being prepared for publication in a scientific journal [20].
- Proposal of a systematic process for performing statistical data pre - processing based on the following processes: perform an imprinting of the main characteristics of a data set and review the main strategies for data cleaning and improvement. The methodology was applied to the Portuguese Surveillance system. This work is being prepared for publication in a scientific journal [21].
- The statistical models applied to large data sets do not fully capture the reporting process of the Portuguese data set. In particular, the conclusion that the fully marginal Poisson process does not capture the underlying phenomena of under - reporting. This work led to a publication presented in [22].
- The study of the time patterns of Portuguese of reporting delay. It was shown that the reporting patterns can be fairly divided into two groups. There are some fluctuations indicating that the reporting is done in batches. This work led to a publication presented in [23].
- The thesis identifies the main factors influencing reporting delays in the HIV - AIDS cases within the Portuguese Surveillance System. This work led to a publication presented in [24].
- It also describes the reporting delay distribution in continuous time taking into consideration the individual, administrative, organizational and historical context. This work is being prepared for publication in a scientific journal [25].

Besides these contributions, the work presented in this thesis establishes work relations with national and international organizations for disease control such as *Commissão da Luta Contra a SIDA*, *European Centre of Disease Control and Prevention* and the *Public Health of Poland*. Contacts with the elements of *Statistics Netherlands* helped to bring the expertise of the UNECE on statistical business methodology to the scientific field.

In the course of this PhD two contributions on close fields were also developed: one modelling of the incidence of the Portuguese Tuberculosis epidemics and the understanding of the Perceived Quality of Life among first-year health students engaged in problem based learning and traditional teaching model: First-year students of health sciences. The first consisted in modelling the incidence through time clustered by regions with similar patterns of incidence [26]. It is well known that tuberculosis and HIV share several social determinants and this work helped to rise this awareness. The second was the first exploration of the multilevel modelling in the context of perceived quality of life among first-year health students [27].

1.5 Thesis Structure

The remainder of this thesis is organized as follows:

Chapter 2 - Infection Diseases and Surveillance Systems. The chapter overviews the main concepts related to the life cycle of HIV, its geographical distribution and also who

is diagnosed and treated. It also presents a description of how public health surveillance systems are organized, how the data is collected and what is the data flow throughout the system. We list the main possible sources of errors that bias the results that arise from it.

Chapter 3 - Statistical Data Pre – Processing. This chapter proposes a data quality methodology for detecting and correcting the main errors. We describe the data sources used in the thesis according to the variables, types and validation rules.

Chapter 4 - Mathematical Models for Reporting Delay Estimation. This chapter explores the mathematical methods used to solve our main objectives. Also, it presents a literature review regarding the modelling, direct or indirect, of the reporting delay pointing the strengths and weaknesses of each method.

Chapter 5 – Methodology and Results. The chapter describes the methodology and the results, showing the mathematical models with increasing layers of complexity for capturing the underlying phenomenon of reporting delay in a surveillance system. It also presents and analyses the results achieved by the methods proposed.

Chapter 6 - Conclusions. The final chapter summarizes the main conclusions and contributions of this thesis and, in addition, some of its limitations and future perspectives.

Software used: The results presented on this thesis were preformed on the following software R, Python, Java and RapidMiner.

Ethical commission: This work was submitted and approved by the ethical commission of Oporto University.

Chapter 2

Infectious Diseases and Surveillance Systems

> *Among the many challenges to health, infectious diseases stand out for their ability to have a profound impact on the human species. Great pandemics and local epidemics alike have influenced the course of wars, determined the fates of nations and empires, and affected the progress of civilization, making infections compelling actors in the drama of human history.*
>
> The Perpetual Challenge of Infectious Diseases, Fauci 2012

Infectious diseases are a result of the penetration of foreign pathogenic micro-organisms into human tissues or organs, followed by their deleterious proliferation in the host [28]. The pathogenic process is generally due to virus or bacteria, fewer times it can also be fungi or parasites. The illness can be spread, directly from one person to another or indirectly from one person to the environment and then to another person, throughout inoculation, airborne or water-borne.

Once the infectious agent enter the body, host's immune system reacts with a response to neutralize the pathogens [29]. This resistance (by prevention or eradication) is mediated by a collection of cells, tissues, organs and molecules. Examples of these mechanisms are the epithelial barriers, lymphoid cells, lymphocytes cells (B and T) and their products such as antibodies [30]. The two major subpopulations of T lymphocytes are the CD4+ and the CD8+ cells. The first regulates immune response upon recognising antigen and the second also have the recognizing ability but also gives rise to cytotoxic T cells [31]. If a harmful pathogen overcomes these defences mechanisms, an inflammatory process is initialized. This process is characterized by a redness, warmth, swelling, pain, loss of appetite, fatigue, and loss of the tissue function [32]. Abnormal immune responses cause many inflammatory diseases with serious morbidity and mortality [30].

Some immunological system cells have the ability to respond more rapidly and effectively to pathogens that have been encountered previously - immune memory, and the vaccines uplift this capacity.

Presently, the top three single agent/disease killers are HIV / AIDS, Tuberculosis (TB) and Malaria, and while the number of deaths due to nearly every disease have decreased, the ones due to HIV/ AIDS have increased fourfold [33].

In our deeply interconnected world of people and goods, infectious diseases constitute a serious threat. An active vigilance for signs of an outbreak, rapid recognition of its presence, diagnosis of its microbial cause is required [1]. It is also necessary to identify modes of transmission and at-risk population groups, and to define public measures for target prevention [2, 3]. The collection of adequate surveillance data is vital to evaluate the disease

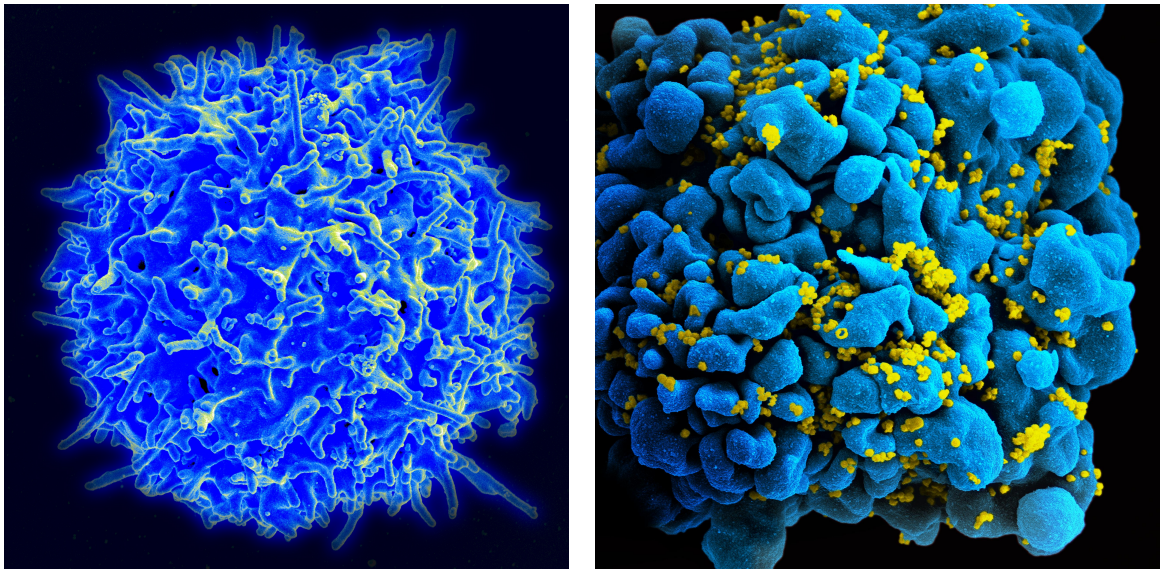


Figure 2.1: Healthy T cell (left) and HIV infected T cell (right)

burden and the impact of prevention and control programmes, aiming for effective and efficient responses.

This chapter provides, in section 2.1, an overview of the HIV life cycle, transmission modes, infection diagnosis and treatment; section 2.2 overviews these same topics about tuberculosis. In section 2.3 described the HIV and TB co-infection dynamics. The epidemiology of the diseases is presented in 2.4. A holistic framework of the epidemic is presented in 2.5. Section 2.6 focus on the data collection process and its main issues are reporting delay, data duplication, completeness and consistency. Section 2.7 discusses these issues for the European reporting processes and overviews the European and the Portuguese Surveillance Systems.

2.1 Human Immunodeficiency Virus - Acquired Immunodeficiency Syndrome

HIV is a biological agent that infects several types of human body cells with affinity for the cells of the immune system, particularly the T cells, that bear the CD4 receptor (CD4+) (Figure 2.1). It is a very complex virus with a rapid replication, high mutation rate and with the ability to escape immune-mediates clearance. Hence, once infection is established, it is never eliminated completely from the body [34, 35]. There are two types of virus: HIV-1 and HIV-2. The first is the predominant worldwide and the last is commonly found in West Africa, and occasionally in East Africa, Europe, Asia and Latin America [36].

Once HIV enters the body, and without any intervention (prophylactic treatment), there is a short term acute illness phase. In this phase the virus is able to integrate into the host's cell's genes and remain hidden from the host immune response initiated 2-4 weeks after the primary infection. People may experience a flu-like illness that may include fever, headache, rash or sore throat. Since they have a large amount of virus in their blood, they are very contagious. Past this few weeks, the HIV virus load drops and begins to proliferate steadily over the following years while CD4+ cell progressively decrease. At this stage there may be an absence of observed disease and this period is the so-called clinical latency or asymptomatic HIV infection because individuals may not experience any symptoms [35, 37, 38]. Without treatment, this stage can last approximately 10 years unless individu-

als are infected with HIV-2; in this case it can be longer. During the viral suppression (when the virus load is at a very low level), the transmission is less likely but not impossible. Over time, a person's viral load increases and the CD4 cell count begins to decrease and so a spectrum of constitutional symptoms start to emerge. Without treatment, individuals start to develop a collection of opportunistic diseases such as constitutional diseases, expedient infections, neurological complications (dementia complex), tuberculosis, lymphomas and Kaposi's Sarcoma (KS) neoplasms that seldom occur in persons with intact immune function [35, 39, 40], and it is said that the person has AIDS. At this stage, the viral load is high and individuals are again very infectious and without treatment; typically it will take 3 years until death.

The typical progression of HIV without intervention is represented in figure 2.2 as well as the decrease of CD4+ cells count. This indicator in normal individuals should range from 500 - 1200 cells / mm^3 and late presentations of HIV infection should be below 350 cells / mm^3 for adults and children with more than 5 years old. Children with AIDS and aged under 12 months-old should have CD4+ cell count under 30 cells / mm^3, from 12 until 35 months-old should have under 25 cells / mm^3 and for children from 36 until 59 months-old should be under 20 cells / mm^3 [37] .

The ratio T4/T8 is often recognized as a quantitative outcome that reflects the critical role of both CD4+ and CD8+ T-cells in HIV-1 pathogenesis or disease progression [41]. The ratio T4/T8 can be approximated by the CD4/CD8 lymphocyte ratio which is approximated 2:1 in normal human peripheral blood [41, 31]. It declines with age but it may be significantly altered by HIV infection.

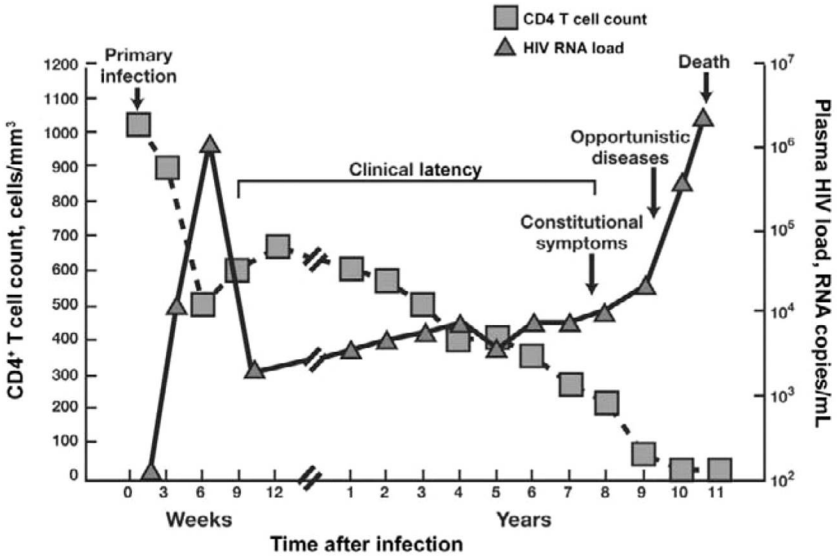


Figure 2.2: Typical progression of HIV infection without intervention (reprinted with permission from Fauci A.S. Pathogenesis of HIV Disease: Opportunities for New Prevention Interventions. Clinical Infectious Diseases, Oxford University Press, 2007, 45, S206-S212)

The diagnosis of the infection is done by the detection of presence or absence of antibodies to HIV in the blood.

The clustering of AIDS cases and the occurrence of cases in diverse groups can be explained by the disease transmission mode. Exchange of body fluids from infected individuals is necessary but not enough. Only certain body fluids - blood, semen, pre - seminal fluid, rectal fluids, vaginal fluids, and breast milk - from infected individuals can spread the virus, both types of virus, although the transmission of HIV-2 is slightly less easy. These fluids must come in contact with a mucous membrane or damaged tissue or be directly injected into the bloodstream (e.g. with a needle or syringe) [35, 42]. So, certain behaviours,

practices and conditions put individuals at greater risk of infection such as:

- having frequently unprotected sexual contact;
- having unprotected anal sex (the highest-risk sexual behaviour) with multiple partners;
- injecting drugs with multi-user unsterilised equipment and drug preparations;
- perinatal transmission from mother to newborn infant;
- receiving unsafe injections, blood transfusions, medical procedures that involve unsterilised cutting or piercing; and
- experiencing accidental needle stick injuries, including among health workers [39, 43, 44].

Although HIV can affect anyone, certain groups are more vulnerable because have one or more of the conditions and behaviour described above:

- Female Sex Workers (FSWs), who have numerous partners, often within a very short period and often can't negotiate the use of condom;
- clients of Sex Workers (SWs), often acting as a bridge to low-risk populations;
- Injecting Drug Users (IDUs), often share injecting equipment with non self blood;
- Men who have Sex with Men (MSMs), because unprotected anal sex has much higher transferability [43, 44].

There is no safe and effective cure for HIV/AIDS but the infection can be controlled. The proper treatment is called Anti-Retroviral Therapy (ART) and it can dramatically increase longevity among infected individuals, reduce the viral load and, therefore, reducing the probability of transmitting to others [**?**, 35]. Although not fully restored, with ART the virus replication rate diminishes and the immune system is capable to protect the individual against some opportunist diseases. So, the proper care may lead to a near-normal, productive life and with a life expectancy that may approach the observed in uninfected population. So, HIV is now considered as a chronic disease [45].

While ART treatment reduces the risk of transmitting, death and complications from the disease, these medications are expensive and have several and serious side effects. Treatment is recommended as soon as the diagnosis is made regardless of the stage [46].

Not only HIV enables opportunistic pathogens that otherwise rarely infect human beings to cause illness, it also substantially worsen the manifestations of other pathogens [47]. TB and Viral Hepatitis are two common co-infections in HIV patients. If individuals have latent TB, with co-infection are more likely to develop active disease and if they have Viral Hepatitis they may rapidly develop liver damage.

2.2 Tuberculosis

TB is the oldest infectious disease known (since pre-historical times) and it is the second most common cause of death from single infectious agent [48]. It is caused by the Mycobacterium tuberculosis (MTB), also known as "Koch's bacillus", and the World Health Organization (WHO) estimates that as much as one-third of the world's population is or was infected by it [49].

Once the micro-organism enters, the host's immune system tries to combat the disease and it is believed that much of the tissue damage encountered has its origin from this response. In the process, CD4 cells are thought to have a central role in the regulation of the immune response to MTB controlling the bacteria growth [49, 50].

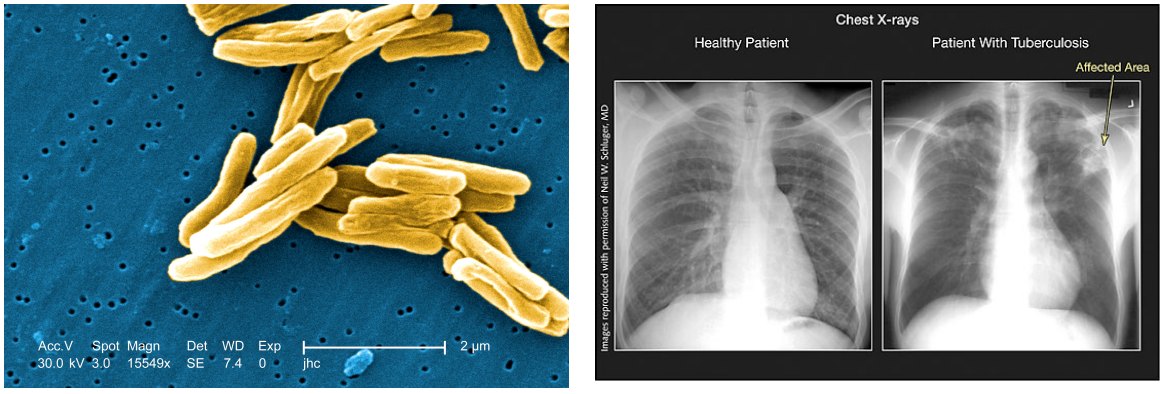


Figure 2.3: *Mycobacterium* tuberculosis bacteria (left) and Healthy and TB infected patient x-ray (right)

In immuno - competent individuals, a relatively small proportion (5 - 15%) of the estimated 2 - 3 billion people infected with MTB will develop TB disease during their lifetime [51]. Commonly, TB occurs in the lungs (90% of the cases) but the bacteria can migrate to other body parts. However, just the pulmonary presentation is infectious (pulmonary tuberculosis) [52]. Therefore, the transmission is done mainly through air from individuals with active pulmonary and/or throat tuberculosis when they cough, sneeze, speak or sing [53]. These infectious droplets can be inhaled by an individual that can become infected or re-infected. With a proper environment, frequent and prolonged interactions, a person with untreated active TB can infect 10 to 15 people each year [40].

It may take many months from the time of exposure to the bacteria until physical symptoms of the infection develop. It has a variety of general symptom such as fatigue, fever, loss of appetite, etc. In particular, on pulmonary form, individuals can experience cough for more than three weeks, cough up sputum, cough up blood, shortness of breath and chest pain. In extra-pulmonary form, the symptoms will depend on what part of the body is involved [54].

The most common method for diagnosing TB worldwide remains on century-old sputum smear microscopy, in which bacteria may be observed under a microscope. In the last few years it is increasing the use of rapid molecular tests to diagnose TB and drug-resistant TB. In countries with more developed laboratory capacity, cases of TB are also diagnosed via culture methods (the current reference standard) [55].

The currently recommended treatment for new cases of drug-susceptible TB is a six-month regimen of four first-line drugs: isoniazid, rifampicin, ethambutol and pyrazinamide, with treatment success rates of 85% or more for new cases. Treatment for Multidrug-Resistant Tuberculosis (MDR-TB), defined as resistance to isoniazid and rifampicin (the two most powerful anti-TB drugs) is longer, and requires more expensive and toxic drugs. The current regimens recommended by WHO, proposes a 20 months' regimen with much lower success rate [55].

Although curable, TB is still a serious threat, having high morbidity and mortality death rate due to the complex treatment regime and to the increasing emerge of drug resistance strains.

The only available vaccine, *Mycobacterium bovis* Bacillus Calmette-Guérin (BCG), is delivered mainly on infants and confers highly variable efficacy against pulmonary disease [56]. In developed countries, and following a continued decline in TB rates, the immunisation is given to children at most risk of exposure.

2.3 HIV-AIDS / TB Co-Infection

HIV and TB have been closely linked since the emergence of AIDS, and both have profound effects on the immune system, as they are capable of disarming the host's immune responses through mechanisms that are not fully understood [57, 58].

HIV increases susceptibility to MTB: not only does the HIV increase the risk of reactivating latent MTB infection, it also increases the risk of rapid TB progression soon after infection or reinfection with MTB [59, 36]. People living with HIV are 29 times (26 - 31) more likely to develop TB disease as people without HIV and living in the same country (exposed to the same risk of TB) [60]. The increased of HIV-associated TB also increases MTB transmission rates at the community level [59]. More over, most of the risk factors associated to one infection predispose the individuals to the other and, therefore, the two infections are concentrated in the same sub-populations [61]. Also, the incidence of TB among HIV infected individuals depends on the number of cases of TB in the country.

TB is the most common opportunistic infection and is a leading cause of death among people living with HIV, accounting for 1 in 5 HIV-related deaths globally. As a corollary, in 2013, 1 in 4 TB deaths were associated with HIV.

In 2014, an estimated 2.0 million persons were newly infected with HIV at a global level and an estimated 9.0 million people developed TB worldwide. Of these, 1.1 million (12%) had both HIV and TB, and 360 000 died from HIV-associated TB [62].

In Portugal, in 2014 and for the first time, the numbers of TB cases were lower than 20 new infections per 100 000 residents. However, the number of HIV infected individuals among TB cases are twice higher than in the European Union [26, 63]. It was shown that the HIV notification rate had a positive and significant effect on TB notification rate. If the rate of notification of HIV increases, it is expected an increase in the TB notification rate. More precisely, an increase of 10 HIV notifications cases per 100,000 population leads to a rise of 2.5 TB notifications per 100,000 population ($p < 0.05$), for a fixed year, geographical location and population density value [26]. So, the Portuguese Direcção - Geral da Saúde started to preventively treat for TB risk groups, such as individuals with HIV - AIDS [63].

There is a strong spatio-temporal association between the two epidemics - Figure 2.4. Evidences show that HIV prevalence is exacerbating the TB epidemic even in regions with low incidence [26] . The WHO has recommended the package of interventions collectively called collaborative TB - HIV activities since 2004 [57, 60].

2.4 Epidemiology

HIV continues to be a major global public health issue, having claimed more than 36 million lives so far and with no clear signs of overall decrease [39, 64]. Seven out of ten people living with HIV are in sub-Saharan Africa, where this infection is a leading cause of death among adults, women of child-bearing age and children [64].

In 2015, Joint United Nations Programme on HIV/AIDS (UNAIDS) / WHO estimated that a total of 31.8 million adults and 3.2 million of children (< 15 years old) were living with HIV worldwide. Although the numbers decreased since 2014, there were still 1.9 million adults and 240 000 children newly diagnosed and AIDS had claimed 1.1 million lives [65].

The epidemic trends and patterns vary widely across European countries. In 50 of the 53 countries of the WHO European Region, there was a rate of 17.6 newly diagnosed infections per 100 000 population which corresponds to 153 407 new HIV / AIDS cases, with the highest prevalence presented in the East of the Region and lowest in the Centre [1](Figure 2.5).

[1]**WHO European Region and Lichtenstein**: **West**: Andorra, Austria, Belgium, Denmark, Finland, France,

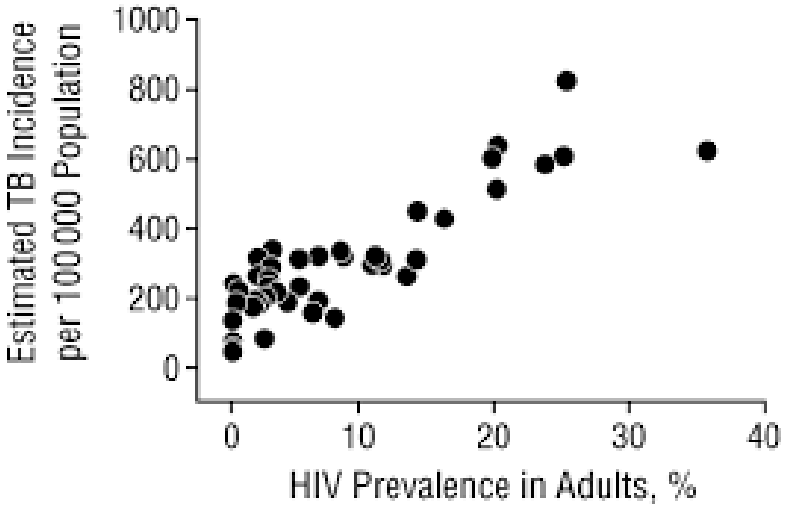


Figure 2.4: Estimated tuberculosis (TB) incidence in relation to human immunodeficiency virus (HIV) prevalence for 42 countries in the World Health Organization African Region [59].

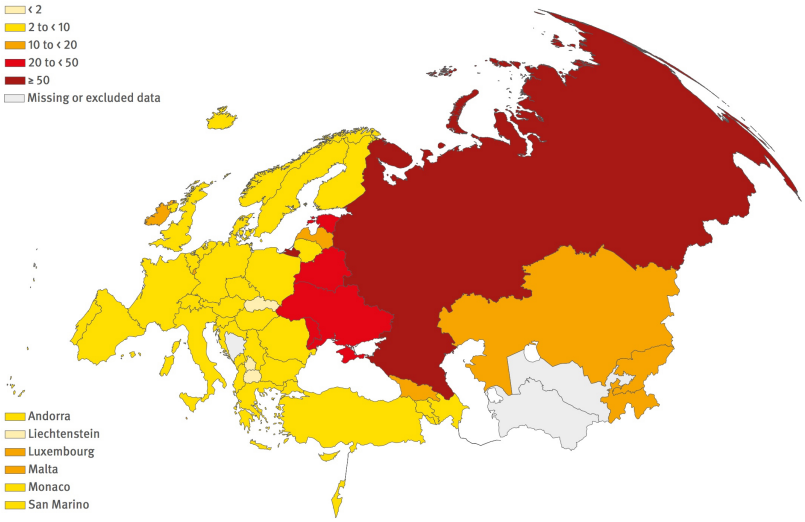


Figure 2.5: New HIV diagnoses per 100000 population, 2015 [66]

The rate of new HIV diagnoses was higher among men being the overall male-to-female ratio of 3.3 and the highest crude age-specific rate was observed among subjects aged 25-to-29 years old.

An excessively high number of people (48%) were diagnosed in advanced stage (CD4 count $< 350 cells/mm^3$). The highest proportion was observed among IDUs (58%) and those acquiring HIV through heterosexual contact (57%), and the lowest in people who acquired HIV through sex between men (37%). The proportion of late diagnoses increased with age, being this characteristic present in 63% of persons aged 50 or older. Being migrant from sub-Saharan Africa and south-east Asia placed people at-higher risk of late diagnoses (both 56%).

Although the number of AIDS cases has consistently declined since the mid-1990s, in 2015 this stage was diagnosed in 3 754 people of 29 European Union (EU)/European Economic Area (EEA) countries, corresponding to a rate of 0.8 cases per 100 000 individuals. The highest rates were reported by Latvia (6.6) and Portugal (2.3). The most common AIDS - defining illness reported by 19 countries was TB (pulmonary and/or extra-pulmonary) [66].

The reported main transmission mode in EU / EEA was sexual contact between men, but in East of the Region heterosexual contact and injecting drug use prevailed [66, 67].

The HIV epidemics of the European Regions in key populations at high risk are intersecting epidemics, in which sexual risks intersect with those related to injecting drug use [68]. SWs involved in injecting drug use have higher HIV prevalence than SWs who do not inject drugs. The first are also more vulnerable to violence, reduced condom use, increased mental health problems and unwanted pregnancies. Studies on risk factor showed a strong and consistent association between an increased number of people imprisoned and increased HIV prevalence among IDUs and FSWs. Prison, an effect of criminalisation of drug use and sex work, is a risk environment for the transmission of HIV [68].

Comparing a 5-year averages of HIV cases according to the reported probable transmission mode and considering that:

1. the exposure group ‘heterosexual’ includes sex work-related transmission
2. risk practices overlap such as some sex workers injecting drugs and
3. exists variation in frequencies of reported HIV cases within each sub-region

data suggest that the epidemic is concentrated among MSMs, IDUs in the West which counts also with the contribution of the heterosexual transmission; although with lower numbers, the Centre epidemic is still concentrated among MSMs, the higher transmission in European Regions occur in the East and it is concentrated among IDUs and with strong contribution of heterosexual transmission [68, 69].

For better understanding the epidemiology of TB/HIV co-infection in the EU / EEA, European Centre of Disease Control and Prevention (ECDC) performed a comparative descriptive analysis of the TB (both HIV-positive and HIV-negative) and AIDS (both with TB-positive and TB-Negative) case-based surveillances data.

The TB perspective In 2014, 21 243 (64.6%) of 32 892 TB cases were reported to have undergone HIV testing, and 1051 (4.9%) of those with known HIV status were reported as HIV-positive [62]. These data were reported by 21 of the 31 EU / EEA countries.

Germany, Greece, Iceland, Ireland, Israel, Italy, Liechtenstein, Luxembourg, Malta, Monaco, The Netherlands, Norway, Portugal, San Marino, Spain, Sweden, Switzerland, United Kingdom **Centre**: Albania, Bosnia and Herzegovina, Bulgaria, Croatia, Cyprus, Czech Republic, Hungary, Montenegro, Poland, Romania, Serbia, Slovakia, Slovenia, the former Yugoslav Republic of Macedonia, Turkey **East**: Armenia, Azerbaijan, Belarus, Estonia, Georgia, Kazakhstan, Kyrgyzstan, Latvia, Lithuania, Republic of Moldova, Russian Federation, Tajikistan, Turkmenistan, Ukraine, Uzbekistan

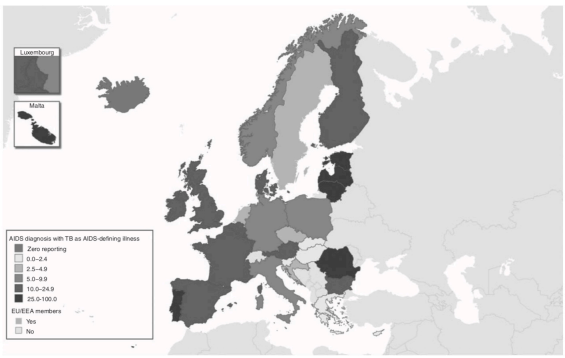


Figure 2.6: Percentage of AIDS diagnoses with tuberculosis reported as AIDS - defining illness, by country, European Union and European Economic Area, 2014 [62]

Figure 2.7: Percentage of HIV-positive cases among tuberculosis cases with known HIV status, by country, European Union and European Economic Area, 2014 [62]

Four countries had HIV testing coverage less than 50%: Czech Republic (28.4%),Denmark (1.9%), Hungary (3.5%) and Ireland (26.6%). Among the 17 countries with at least 50% reporting completeness, the proportion of co-infected cases was highest in Latvia (19.5%) followed by Malta (17.1%), Portugal (14.7%) and Estonia (10.1%) 2.6 [62].

The AIDS perspective In 2014, 3863 cases of AIDS were reported by 29 countries in the EU/EEA. TB was the second most common AIDS-defining illness with 691 (17.9%) (the first more common was *Pneumocystis* pneumonia). Among these co-infected cases, 497 (71.9%) were reported as pulmonary TB and 194 (28.1%) as extra-pulmonary TB. The highest proportions of TB as an AIDS-defining illness were reported by Malta (75% of AIDS diagnoses), Romania (43.9%), Latvia (41.5%) and Lithuania (37.8%), whereas in Cyprus, Greece, Slovenia and Slovakia no AIDS cases were reported as presenting with TB as AIDS-defining illness 2.7 [62] .

2.5 HIV / AIDS Socio - Ecological Framework

Thirty-five years of HIV / AIDS history led to a clear understanding that reducing its incidence is not simply a battle for using condoms and keeping individuals engaged with medication [70]. The unsuccessful individualistic or single-issue focussed interventions led to a clear recognition that “HIV risk behaviour can only be significantly and sustainably changed by considering all the up-stream factors which shape those patterns of behaviour” [71, 72].

Practices are socially produced behaviours that are organised and patterned by culture [73]. So, HIV / AIDS epidemic is shaped by social and structural drivers as well as by individuals characteristics [74, 72, 68]. These multifaceted factors and their relationships can be explained by a Socio-Ecological Model (SEM) which is based on the premise that while individual levels risks are necessary for the disease spread, they are insufficient to explain population level epidemic dynamics [72] . This model typically contains five hierarchical levels highly interactive throughout mechanisms of critical importance:

1. individual,
2. interpersonal / network,
3. community,
4. organizational / institutional / health system, and
5. structural / policy / environment.

The *individual level* corresponds to biological or behavioural characteristics associated with vulnerability to acquire or transmit illness or infection [72]. It includes factors such as perceptions, beliefs, emotions, knowledge, attitudes, self-efficacy, developmental history, gender, age, religious and racial/ethnic identity, sexual orientation, economic status, financial resources, values, goals, expectations, personal stigma, fear of others stigma among others [70, 71, 73, 72, 75].

The *interpersonal /network* level consists of interpersonal relationships including family, friends, peers, co-workers and other that directly influence health and health behaviours in multiple ways [72]. Includes factors such as relationship satisfaction, social support systems, customs or traditions and networks, which are not bounded by geography, socioeconomic status, or cultural and racial [70, 71, 73, 72, 75].

The definition of what is a *community* is not consensual but generally includes: network ties, relationships between organizations and groups, and geographical / political regions [72]. This level includes influences such as gender norms, stigma, mobility of the local population, social capital, community norms, informational networks within defined boundaries, the built environment (e.g., parks), village associations, community leaders, businesses, and transportation.

The *organizational / institutional* level focus on factors within the health system with rules and regulations for operations that affect how, or how well services are provided. Examples of these factors are quality of services, confidentiality, sufficient resources, support tools, convenient and responsive services.

At last, the *structural* can be described as a level of local, state, national and global laws, legal rights or regulations and polices, including polices regarding the allocation of resources[75]. This last macro-level contains factors that are out of individuals control, acting as barriers or facilitators, mediating lower order risks such as those at the individual or interpersonal / network levels [72]. It includes factors affecting behaviour, such as economy, political climate, enforcement of policies and laws, or funding dynamics, poverty, educational curriculum and gender equality [70, 71, 73, 72, 40].

Figure 2.8 represents several factors at each level of the SEM for HIV-related behaviour and /or behaviour change; although over simplified it is consistent with the majority of the ecological literature [70]. Note that factors can span levels and therefore the boundaries between them may be understood as permeable, also the relative importance of each factor varies with population and local context [72]. For example, the behaviour - condom use - has varied social and cultural meanings depending whether it concerns a 'marriage bed', a brothel or casual sexual encounter [73].

Individual characteristics such as knowledge and skills may influence the perception about what may be a HIV risk behaviours but may also affects the capability of taking the

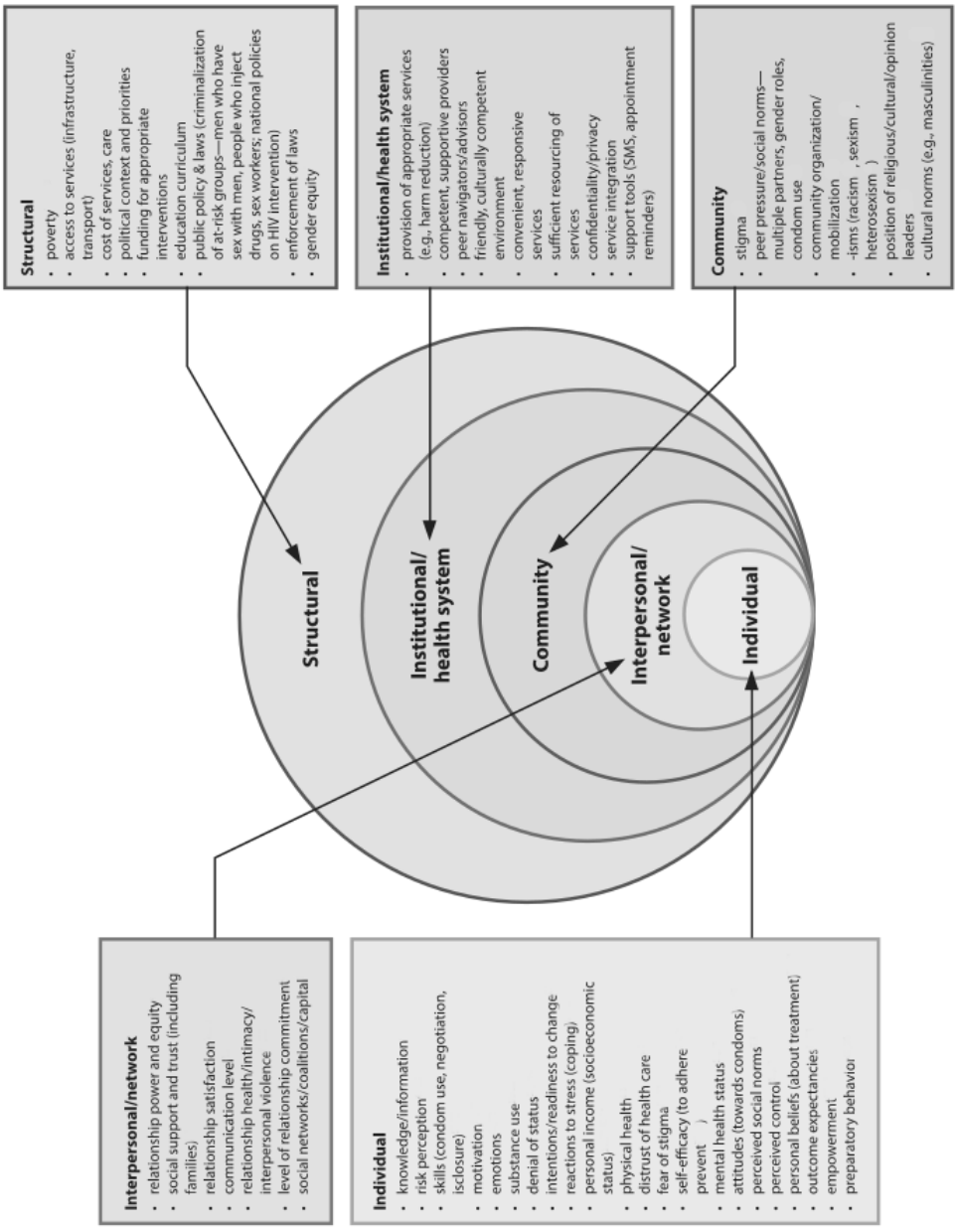


Figure 2.8: Factors influencing HIV-related behaviour and / or behaviour change at each level of the socio-ecological model [70] (reprinted with permission from Michelle Kaufman)

necessary actions to self - preservation [74]. For example, individuals that have low health literacy concerning HIV-prevention methods are more vulnerable, thus being at a higher

risk of becoming ill. Another example is low self-esteem that may lead individuals to seek multiple sexual encounters in an attempt to self-validation, or lead individuals taking refuge in the use of alcohol or drugs — substances that impair judgement and may interfere with the risk perception and with the person's ability to negotiate and practice safer sex — to enable sexual encounters [76]. Individuals in both high and low-risk contexts can reduce their sexual risk of contracting HIV- by being abstinent, selecting HIV-uninfected partners, being monogamous with an HIV-uninfected partner, engaging in safer sex practices (e.g., always using a condom), and/or adhering to a pharmaceutical pre-exposure prophylaxis regimen [77].

Individuals cannot exist without social networks which directly influence behaviours, even when the individual may wish to act differently [70]. A well-known protective factor is the social support and trust throughout family and social networks, reinforcing social norms and behaviour. On the contrary, sexual and social networks with high infection rates and high prevalence of risk behaviour such as multiple sexual contacts and share of used injection drugs paraphernalia exponentiates HIV transmission between the members of the network [72]. Other barrier is an unfriendly social and cultural environment that can create stress and thus interfere with the individuals' ability to self - health promotion.

History provides an unfortunate abundance of examples of community prejudice, discrediting, discrimination, discounting and mistrust directed toward particular groups of individuals [76, 78, 74]. It is widely recognized that these phenomenons are a persistent and pernicious barrier to prevention, uptake of HIV test and treatment creating health risk and worsening health care [70, 79, 80, 81]. They may also be an obstacle to the proper communication of the transmission modes, and may induce individuals to deny their serological status and neglect their specific needs [78].

Two out of three countries in Europe and Central Asia acknowledge that stigma and discrimination within key HIV affected populations are a barrier to epidemic control and contributes to late diagnosis. These negative attitudes among health professionals, particularly with respect to sex workers, men who have sex with men and people who inject drugs, reportedly persists across the region and plays a role in preventing these key populations from accessing HIV prevention, testing and treatment [81].

Other important driver factors are the interpretation of community norms and public HIV-positioning of influential members of the community affecting the provision and/or uptake of HIV prevention, testing, treatment and access to care services [76, 72].

The deprivation of convenient, competent and responsive health resources is a clear barrier to epidemic control. Among these resources, health providers are essential given that they are the first contact with patients who need counselling, support and proper treatment. So, effective and supportive providers are determinant to patients trust in their counselling and following their recommendations. Confidentiality and privacy of health institutions are other key elements due to stigma and illegality of some HIV-related behaviours.

People living with HIV and AIDS need a variety of healthcare and social support, because of their vulnerability to opportunistic infections and their progressive disease [82]. A special and essential role of counselling, prevent, monitor the evolution, support and comfort are provided by the Physicians' which have a privileged point of global attention. The global shortage of health workers is estimated at over 4 million, with 57 countries experiencing severe shortage; defined as fewer than 0.23 doctors, nurses, or midwives per ten thousand people. This translates into nearly a billion people who have no access to a health worker of any kind of the one- seventh of the world's population. The majority of identified HIV-positive people are in contact with health care providers. [82]

Laws criminalizing injecting drugs, same-sex behaviour or commercial sex may cause an individual feel fear of discrimination or even of being arrested when seeking a health care institution for diagnostic and treatment. Also, may have an impact on stigma and mis-

conceptions at the community and interpersonal levels affecting the adoption of prevention and treatment behaviours [70]. Another example is the economic inequalities at the macro-social scale that can divide communities, encourage transactional sex and introduce vast interpersonal power inequalities in the negotiation of safer sex, affect the access to forms of prevention such as buying condoms and prophylactic treatment [70, 83]. Countries economy affect the quality of health care, quality of intervention programmes and the individuals' acceptability of the services which "is a concept that is embodied in individuals and is thus affected by perceptions of the accessibility of the healthcare, health system responsiveness, and individuals psychological status, experiences and expectations" [83]. This ultimately impacts satisfaction, adherence to prescriptions and treatment outcomes which increases the risk of poor health conditions, onward HIV transmission and ultimately death [83]. Policies determine allocation of economic resources to education, health care, job training, financial assistance and HIV prevention services and therefore play a substantial role in shaping structural contexts of HIV risk [72].

More than one in three countries in Europe and Central Asia report that unfavourable laws and policies are a barrier to provision of HIV prevention services for people who inject drugs and sex workers. In two out of five countries, such laws and policies reportedly limit the provision of HIV prevention services, including harm reduction intervention, in prisons and in half of the countries, laws and policies are reported to limit access to treatment for undocumented migrants [84].

Knowledge about disease incidence is vital to its control so a proper case register system must be implemented in each country.

2.6 Public Health Surveillance Systems

Surveillance has been defined as the continual scrutiny of all aspects in emerging and spread of a disease that are pertinent to effective control, involving a systematic collection, analysis, interpretation, and dissemination of health data [1, 12, 2]. Surveillance systems rely on processes and individuals, thereby can be found to differ substantially according to the disease, health condition and country. Nevertheless, all traditional public health surveillance approaches use pre-specified case definitions, employ manual data collection and human decision-making [4, 3, 5, 6].

For surveillance purposes, before counting cases it is necessary to decide what a "case" is. Case definitions are a set of standard clinical and laboratory criteria that unequivocally classify whether a person has a particular disease, syndrome or other health condition developed by epidemiologists [85, 3]. These case definitions remove the potential bias and make the comparison between populations possible in different geographical locations and times. So, they are a fundamental cornerstone for standardising the collection of data [3].

Typically, epidemiologists in national public health authorities collect confirmed cases from laboratories, general practitioners and hospitals which are common detection and primary infection diagnosis data sources [3]. Each case is record, processed and compiled in a national database.Then the information is analysed and made available to public health professionals, the general public and public health authorities for supporting decision-making in public health practices [15, 3].

In Europe and in accordance with the ECDC founding regulation (Regulation (EC) 851/2004), all EU Member States (28) have to provide to this centre in a timely manner, the available scientific and technical data on 52 communicable diseases and related special health [86, 87], using The European Surveillance System (TESSy) which is hosted and validated by ECDC. In the case of HIV, TB and Influenza, the surveillance is conducted with the collaboration of the World Health Organization - Regional Office for Europe (WHOE) (Figure 2.9).

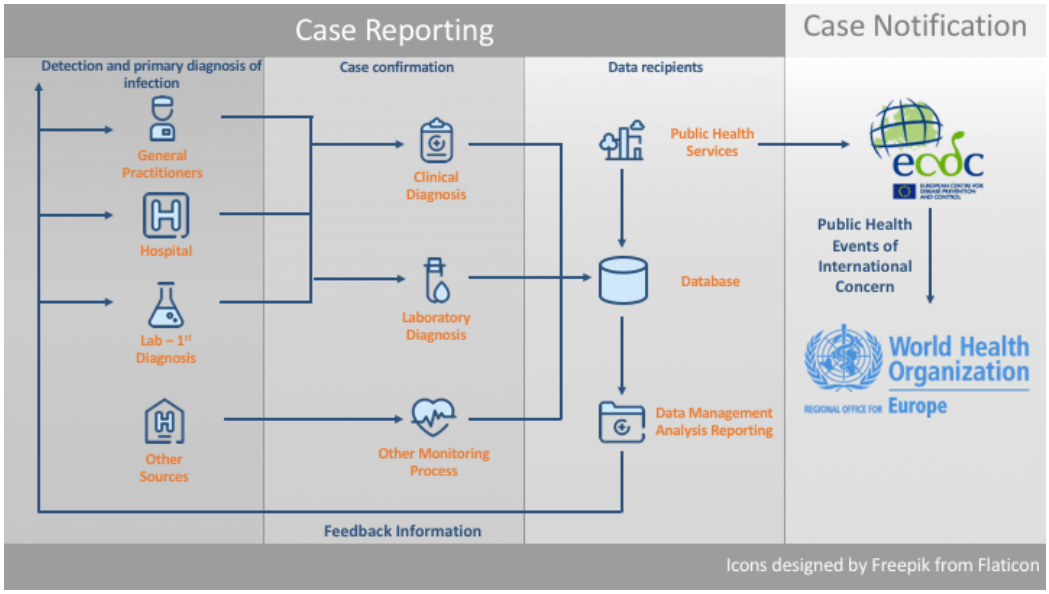


Figure 2.9: Simplified flowchart of a generic surveillance system based on [3] and [88]

Considering the attributes of the disease and objectives for which they were implemented, surveillance systems may be classified as passive or active, as compulsory or voluntary and as comprehensive or sentinel. In a passive system, the data providers take the initiative to report the case to the public health agency which do not stimulate the reporting nor give feedback. By the contrary, an active surveillance relies on the initiative of public health official that contact relevant sources of data, stimulating them to report, send agencies alert or remind and give feedback of the results [3, 89]. Some systems make data submission mandatory by law, professional edict, policy or guidance. Comprehensive systems include reports of cases that occur within the whole population of a geographical area covered by surveillance system while sentinel rely on notifications from a selected group of reporting sources [3, 89]. Active and mandatory systems typically generates high-quality data (defined as high levels of completeness, validity and timeliness).

The information reported can be aggregated or case-based. The latter occur when individual anonymized data is reported while in aggregate-based systems only the information related to a group of people classified under the same category for a disease or health-related event (for example the number of people reported by age and / or risk group) is registered [3].

The format in which data is collected may be paper or electronic through an information system. The paper format has the advantage of requiring no special technical skill and it entirely circumvents the difficulties of interfacing between information systems and the disadvantage of needing to undergo a transcription step increasing the risks of introducing additional errors and do not enforce the completion of mandatory fields or other validation checks [3].

Moreover, given the typical fragmentation of surveillance systems and the demands of health care systems, several problems inherent to data must be recognized. This issues may occur in each and every step of the data collecting process, data management and analysis, affecting negatively the aim for which the data was collected [12, 13, 3, 14, 15, 16]. The most common problems are unrepresentative or selection bias , changes in the system implementation through time, inconsistent case definitions, availability of cases, reporting delay, and processing errors such as errors during completion of the reporting form or data entry.

The first problem causes sampling errors on data obtained from the surveillance reg-

istries while the others may introduce bias on statistical estimates commonly designated by non-sampling errors. Changes in the system implementation through time are inevitable. As epidemiologists' knowledge about the natural history of the disease being surveyed and diagnosis techniques evolve its natural that the case definition change. Likewise, the evolution and adopting of different information technologies used for reporting a case may affect the behaviour of system main stakeholders. Other important factors affecting data quality are changes in government funding, legislation or reporting sources. It must be recognized that any change in the reporting system may interfere with the number of cases recorded and thus with the statistical analysis, especially in time trend monitoring. Moreover, time changes in surveillance infrastructure, clinical practices and reimbursement policies may smite representativeness [3].

2.6.1 Surveillance Data Representativeness

Generally, health conditions are not reported randomly. Diseases handled in a public health facility are reported disproportionately more frequently than those diagnosed by private practitioners; conditions that lead to hospitalization, are more likely to be reported than problems dealt with on an outpatient basis; diseases with testing practices implemented by the central government are more likely to be diagnosed and reported [12, 3, 85]. So, the information collected may not be representative of the affected population and thus the occurrence of the health-related event over time and its distribution in the population by place and person may not be accurately described [3]. Collecting data from multiple sources may help to provide ways to improve the representativeness of the information.

The unregistered cases may occur at a community level, when patients do not seek professional care, in a surveillance system level when the system fails/misses or delays the reporting of diagnosed case or / and in the health department itself when, for example, cases are lost due to misclassification [90, 3, 7]. A general morbidity surveillance pyramid, is often used to illustrate the availability of disease data at each surveillance level. "With each ascending level (from the community, to healthcare institutions, to regional and national public health agencies) data availability shrinks and only a fraction of cases from the level below is captured. In contrast to the narrow tip of the pyramid which represents data held by national public health agencies, the base is wide as it holds all infections in the community. The difference between the numbers at the top and at the base can be considered cases lost to underestimation" which is the sum of cases lost to under-ascertainment and cases lost to underreporting. The typical morbidity pyramid is presented in Figure 2.10 [90]. Depending on the disease, 5% to 80% of the cases that actually occur will be reported [12, 3].

Under-ascertainment or under-diagnose of cases correspond to patients that are not diagnosed and hence not identified by the healthcare systems and it may occur when:

- patients do not visit the healthcare services because they do not feel symptoms or feel only mild symptoms
- a patient have low health literacy and do not recognize symptoms nor perceive the need to seek healthcare;
- unequal geographical distribution of healthcare services exists;
- diagnostic tools are unavailable;
- routine surveillance does not capture marginalised high-risk groups (e.g. commercial sex workers, IDUs, MSMs);
- physicians and the general population are unaware of the disease [3].

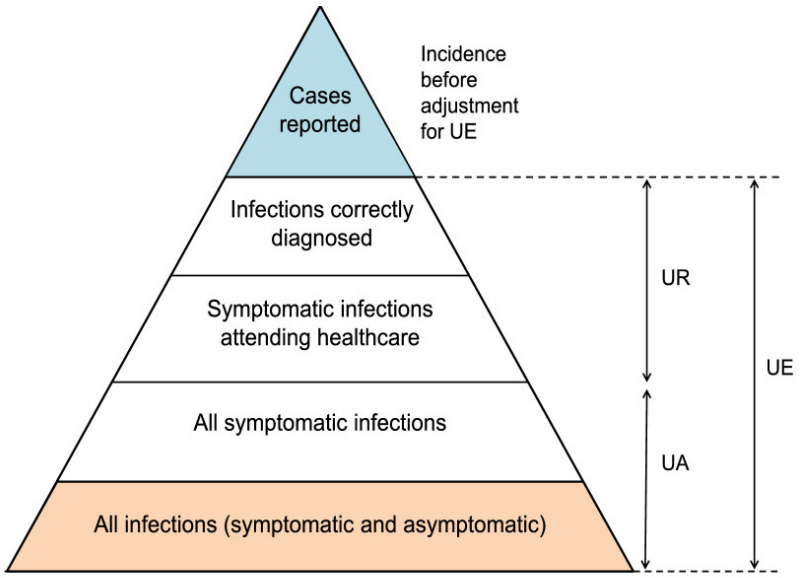


Figure 2.10: Typical morbidity surveillance system [90].
UR - underreporting, UA-under-ascertainment, UE - Underestimation

Underreporting of cases occurs when patients visit physicians or healthcare facilities, but once diagnosed, they are not properly reported to the local public health unit [3, 12]. As ECDC points, it may occur when:

- patients during visiting the doctor do not show symptoms of the notifiable disease;
- a patient has more than one reportable condition and only one is reported (e.g. AIDS with TB as opportunistic disease);
- the surveillance system (e.g. communication system or software) does not fully support the actions initiated by the healthcare provider (or reporter) so he/she cannot file the formal report;
- an attempt to a report was made but the communication or the software system failed;
- the patient is presenting a serious clinical condition and the physician focuses on the patient, simply forgetting to notify the case;
- the registered case is present and registered in the public health surveillance database, but has been misdiagnosed, misclassified, or miscoded [3].

Assessing the non-visible cases (under-reported or under-diagnosed) is a difficult task [3]. When the disease is on a symptomatic stage other sources of information may be useful, such as:

- monitoring sales of specific drugs over time through electronic systems;
- investigating the symptomatic cases social networks for identification of non-visible cases;
- screening of risk groups or samples of the general population (e.g. pregnant women, refugees, prisoners, sex workers, blood donors, pre-surgical patients). [3]

When the disease is on an asymptomatic stage the only way to find the patients would be through specific sero-surveys (e.g. serological surveys). This could be relevant for diseases such as hepatitis B/C and HIV [3].

2.6.2 Reporting Delay

Reporting delay can be defined as the time between the diagnosis of a case (by the physician and/or the laboratory) and its report to the national surveillance system. Timely report may enable public health authorities to take fast and effective actions to prevent outbreaks by reducing disease transmission in a population [91, 92, 93]. Reporting delays may depend on a number of factors such as:

- the use of confirmatory laboratory testing;
- the volume of cases identified in the reporting source;
- the reporting process by the health care provider or the laboratory to the local, region, or state public health authority;
- the case follow-up investigations to verify the case report or to collect additional case information;
- periods of decreased surveillance system activity due to reduction of financial resources;
- computer system down-time for maintenance, upgrades, or new application development;
- data processing routines, such as data validation or error checking;
- physicians workload;
- different geographical location [5, 8];
- reporting format.

2.6.3 Form Completion or Data Entry Errors

During completion of the reporting form or data entry, different types of errors may occur:

- Errors in interpretation or coding, for example:
  - Unclear case definitions or that are not widely known may lead to variation on the criteria use. The greater complexity of the diagnosis, “the greater the difficulty in reaching consensus on a case definition” and properly register the case. For example, pediatricians sometimes report diagnosed cases of childhood diseases such as measles as verified, even though the EU case definition requires laboratory confirmation;
  - frequently adjustments to the disease case definition due to evolution of disease-related knowledge, may introduce a confounding effect;
  - emerging of new diseases or definition of parameters that the surveillance system has not taken into account cannot be reported. When *S. enterica serovar* was first recognised, some systems reported related cases either as unknown or as *S. typhimurium* because this new monophasic Salmonella serovar was not included in the pre-populated list of values [12, 3, 94].
- Errors of intention, for example:
  - selecting only cases with good outcomes to report (“cherry-picking”). Avoidance or detection of intentional error can be challenging. Some approaches include checking for consistency of data between sites, assessing screening log information against other sources (e.g., billing data), and performing onsite audits (including monitoring of source records) either at random or on purpose.

  - when some of the information of the reporting form is self-reported; for example, in diseases like HIV, individuals may not report embarrassing conditions, socially discriminated or criminalizing activities [95];
  - stakeholders perceive negative consequences to their interests by reporting high infection rates, so intentionally report lower ones [96].

- Errors when data is entered into the registry inaccurately, such as incompletion of the reporting form, data entry, transfer, or transformation accuracy. These errors are due to insertion, deletion or substitution of characters or writing wrong dates, such as:

  - a laboratory value of 2.0 entered as 20;
  - it has been documented that data entry operators show a systematic digit preference when filling in numeric fields for age (ages ending in 0 and 5 are overrepresented) or date (01, 10, 15 and 20 are overrepresented);
  - digit preference may be also combined with the tendency to avoid certain 'unpleasant' numbers, such as 13 [3].

The most common errors are random ones, during completion of the reporting form or during data entry, namely insertion of characters, missclassification of the disease, changing code of observations, writing wrong dates and do on [3]

Avoidance or detection of interpretive error can be achieved by adequately training main providers on definitions and testing against standard charts. Avoidance or detection of accuracy errors can be achieved through upfront data quality checks (such as ranges and data validation checks), re-entering samples of data to assess for accuracy (with the percent of data to be sampled depending on the study purpose), and rigorous attention to data cleaning [94].

From Population under Surveillance or Coverage until Central Government, it must be recognized that the surveillance data collection process is composed by several critical levels with unique mechanisms. It has a strong human decision-making component firmly imprinted in data which may affect the information quality registered on the system. From representativeness until simple data entry errors several issues may occur. The main issues described in the previous sections are summarized in Figure 2.11 and grouped by the key levels of the system.

A particular case of a surveillance system is that for the HIV/ AIDS. It differs from other surveillance systems in many ways, namely by reflecting the special transmission patterns, the long asymptomatic or with mild symptoms latency period of the infection, the lack of affordable treatment and cure, high case fatality rates, and the social stigma associated with it [9].

2.7 European Surveillance of HIV/AIDS

Surveillance for HIV/AIDS is among the most complex surveillance systems [13]. Its main purposes are to determine the extent of the epidemic and to track the changes or trends in the epidemic overtime [97].

HIV - AIDS surveillance collects, analyses, and disseminates information about new and existing cases of HIV infection (in all stages of development). Given the long asymptomatic (or with mild symptoms) incubation period between infection and the development of AIDS, a surveillance system that relies solely on AIDS case reporting is not effective since these data are usually used to reconstruct the past prevalence of HIV infection and are not appropriate for tracking current infections [9]. In addition, the definition of AIDS becomes increasingly meaningless with the provision of Highly Active Anti-Retroviral Therapy

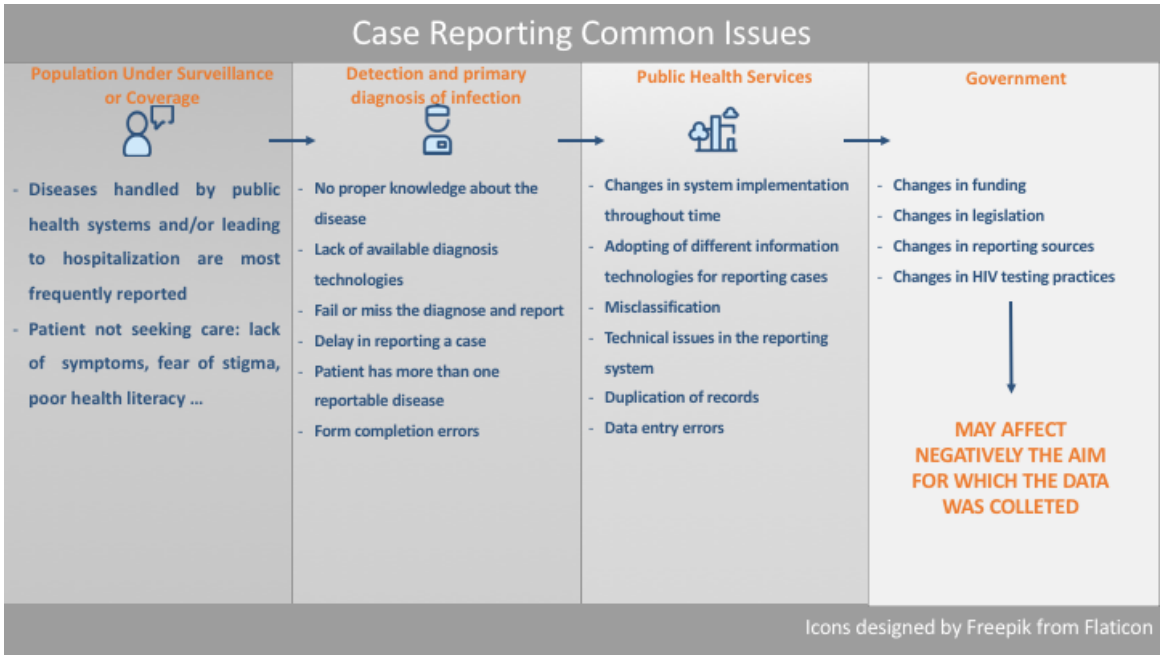


Figure 2.11: Case reporting main issues from a surveillance system

(HAART) [97, 9, 68] since this stage is becoming rarely observed (when patients undergo proper treatment) .

Since 1982, many definitions have been used for national and international reporting. In 2012, the European Parliament and the Council established case definitions for AIDS and HIV on the Commission Decision of 27/09/2012 to the community network under Decision No 2119/98/EC [98]. This definition of AIDS includes all HIV - infected individuals with CD4 counts less than 200 cells $/\mu L$ as well as those with certain HIV - related conditions and symptoms. Although the fine points of the classification system are rarely used in the routine clinical management of HIV infected patients, a working knowledge of the staging criteria is useful in patient care [99, 100]. The details of the definition are presented in Appendixes A and B.

HIV/AIDS surveillance also poses a number of special ethical problems arising mainly from the stigma and discrimination attached to AIDS. Finally, HIV/AIDS surveillance includes behavioural surveillance, in order to understand trends in behavioural risk factors for HIV [9].

All European countries have developed systems to monitor the evolution of AIDS [68]. Since the identification of the first case in the early 1980s, the focus of surveillance was on AIDS reporting. With the development of HIV laboratory technologies and the introduction of HAART the number of AIDS diagnoses no longer reflects the underlying trends in the HIV epidemic satisfactorily [101]. The collection of these cases remains important for providing insights into the extent of late diagnosis and the impact of HIV treatment [68, 101].

All countries in the WHOE, except Monaco and Liechtenstein, have established systems for monitoring the number of new HIV diagnoses [68]. These systems have been replacing the AIDS surveillance to overcome their major limitations: when used to monitoring the HIV epidemic they do not represent the true HIV - incidence, reported cases may include infections that occurred several years previously and are dependent on uptake of HIV testing practices in the population [102, 68].

In some countries, systems have undergone significant revision in the way that they operate; as a result, data from different time periods are not always comparable and some have no national coverage [68].

Surveillance systems are not identical across Europe. In addition to the surveillance data common quality issues discussed in previous sections (in particular, reporting delay, case duplication,completeness,...) , one must recognize that countries vary in data collection methods and testing policies that may have an impact on the the results and introduce bias in comparisons between countries[68, 103, 66].

Due to these variations, national data need to be compared carefully since countries with the largest number of diagnosed cases may reflect not the true scale of the epidemic but the efficacy in case finding.

2.7.1 Reporting Delay

It is widely recognized that reporting delay is an ever present issue that downward biases the HIV trend estimates. This bias is greater in most recent years and, to a lesser extent, in the 2 to 3 years prior to the reporting period [66].

In 2006, a European-wide survey of HIV surveillance systems in 44 countries found that among the 16 countries that had examined reporting delay, the majority stated that 90% of the HIV diagnoses were reported within 6 months and only 3 countries stated that 75% of cases were reported in the same time period [68, 102]. In present days, it was estimated that the longest delays occurred in 9 countries: Greece, France, Italy, Luxembourg, the Netherlands, Poland, Portugal, Sweden, and the UK [103].

The key challenge related to the reporting delay are specific one-off actions in surveillance aiming to capture previously unreported cases.These cases are subsequently reported with excessively long delays. The analysis of the trends in reporting delays exhibits peaks related to control activities in surveillance (“cleaning events”) in Portugal (2013 – 2014), the Netherlands (2010 – 2011) and Poland (2010) [103].

Underreporting The exact extent of underreporting is unknown. Fewer than 40% of European countries have evaluated their surveillance systems for this limitation and only two have published the results. Previous estimates states that it ranges from 0% to 25% for AIDS cases, while for HIV it can range from 10% (Iceland and Italy) to around 40% (Germany and the UK) [68, 102, 66].

In many cases, underreporting is viewed as a special case of reporting delay; it is considered a reporting delay with infinity length.

2.7.2 Duplication

The duplication of HIV/AIDS reported cases occurs in several European surveillance systems [66, 103] and may be originated by the following settings:

1. an individual may have more than one positive HIV test as a result of receiving health care in different settings or using both anonymous and named testing services;
2. a single positive test could be reported more than once, for example, by both the laboratory undertaking the testing and the clinician [68];
3. a single case reported by two physicians in two different settings; for example, a positive test on an Injecting Drug User (IDU) that is engaged in a drug treatment program may be reported, simultaneously by the clinic when the treatment occurs and by the general practitioner.

2.7.3 Completeness and Consistency of Risk Factors' Details

Missing data are also a well-recognised problem within surveillance systems that may lead to biased and potentially less precise estimates. In principle, whenever reporting delays or missing data occur, the accuracy of epidemiological distributions and trends is questionable [103].

The completeness of key epidemiological variables such as age, sex, transmission category, CD4 count at diagnosis and migration status (either of the following: country of birth, region of origin, country of nationality) vary over time and between countries [103, 68, 66]. In 2015, ECDC and WHOE reported the lowest completeness rate being observed on CD4 count [66]. This rate increased from 25.9% in the pre-2000 period to 56.8% during 2012-2014, corresponding to the improvement in CD4 availability, which increased from 31.5% among those diagnosed before 2000 to 63.1% for the 2012-2014 period. Some countries did not provide any CD4 counts and others included this variable in surveillance with delay. In countries who included the CD4 counts in the form, the information is systematically missing either at all or for the earlier years. Migration status could be determined for 88% of the cases diagnosed during all analysed time periods after 2000 compared with 66% for those diagnosed pre-2000. On the other hand, the proportion of individuals with known transmission category decreased slightly from 89.2% among those diagnosed pre-2000 to 81.7% for those diagnosed in the period 2012-2014. Age and gender were consistently reported in proportions greater than 98% during all periods [103].

HIV Testing Practices

HIV testing practices have a direct effect on the extent to which HIV infections are diagnosed and reported an hence also on the representativeness (this is called a non-sampling error). Approaches to HIV testing vary widely in the European region, but most countries have a policy or strategy to offer HIV testing and counselling to most-at-risk groups and pregnant women [68, 104]. In 2015, ECDC commissioned an evaluation of the HIV testing practices through a distribution of a survey [104] to a Primary Target Group Member States. The survey showed that, considering their national guidelines, 78% of the countries had post-test access to treatment, care and prevention services, 70% of the countries had voluntary, confidential testing with informed consent and post-test counselling, and 57% reported having a pre-test counselling or pre-test discussion. Dedicated HIV testing centres (e.g., for people at a high risk such as IDUs) are present in 65% of the countries. Testing of all pregnant women for HIV is recommended in 70% of the countries and this practice is offered to all most-at-risk population groups in 57% of the countries [104]. Table 2.1 describes the testing practices and their distribution throughout the countries according to the target sub-groups, its goals and principles and setting.

Table 2.1: Elements included in the HIV testing practice (n=23). Source: ECDC in [104]

Elements	% (n)	Member States
Sub-groups		
Testing of all pregnant women for HIV (opt-out)	70% (16)	Austria, Belgium, Czech Republic, Estonia, Finland, France, Greece, Ireland, Italy, Lithuania, Malta, Netherlands, Norway, Portugal, Romania, Slovakia
Testing of all most-at-risk population groups	57% (13)	Belgium, Croatia, Czech Republic, Estonia, Finland, France, Greece, Italy, Lithuania, Netherlands, Norway, Portugal, Romania
Testing of sero-discordant couples (routine testing of HIV negative partner)	52% (12)	Croatia, Czech Republic, Estonia, Finland, France, Greece, Lithuania, Malta, Netherlands, Norway, Portugal, Romania
Testing of people with an indicator condition, e.g. pneumonia, mononucleosis-like illness	48% (11)	Belgium, Croatia, Estonia, Finland, France, Greece, Italy, Lithuania, Netherlands, Portugal, Romania

Table continues...

Table 2.1: Elements included in the HIV testing practice (n=23). Source: ECDC in [104] - continue

Elements	% (n)	Member States
Testing frequency	30% (7)	Estonia, France, Greece, Malta, Norway, Portugal, Romania
Tandem Hep C or B/C and HIV testing	22% (5)	Estonia, Finland, Greece, Malta, Romania
Goals and principles		
Post-test access to treatment, care and prevention services	78% (17)	Belgium, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Greece, Ireland, Italy, Lithuania, Malta, Netherlands, Norway, Portugal, Romania, Slovakia
Voluntary, confidential testing with informed consent	70% (16)	Belgium, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Greece, Italy, Lithuania, Malta, Netherlands, Norway, Portugal, Romania, Slovakia
Post-test counselling	70% (16)	Belgium, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Greece, Italy, Lithuania, Malta, Netherlands, Norway, Portugal, Romania, Slovakia
A defined target audience	65% (15)	Belgium, Croatia, Czech Republic, Denmark, Estonia, Finland, Greece, Ireland, Italy, Lithuania, Netherlands, Norway, Portugal, Romania, Slovakia
Pre-test counselling or pre-test discussion	57% (13)	Belgium, Croatia, Czech Republic, Estonia, Finland, Greece, Italy, Lithuania, Malta, the Netherlands, Norway, Portugal, Romania
Desirability or requirement to remove legal or financial barriers	43% (10)	Belgium, Czech Republic, Croatia, Denmark, France, Greece, Italy, Norway, Portugal, Slovakia
Raise professional awareness and train the workforce	43% (10)	Croatia, Czech Republic, Denmark, France, Greece, Italy, Lithuania, Malta, Portugal, Romania
Monitoring and evaluation programme	39% (9)	Croatia, Czech Republic, Denmark, Finland, France, Greece, Netherlands, Portugal, Romania
Partner notification	35% (8)	Czech Republic, Denmark, Estonia, Greece, Lithuania, Malta, Netherlands, Norway
Testing conducted by lay providers	22% (5)	Czech Republic, Denmark, France, Greece, Portugal
Streamlining of counselling process (less demanding)	17% (4)	Croatia, Denmark, Finland, Greece
Written informed consent	9% (2)	Lithuania, Romania
Settings		
Dedicated HIV testing centres (e.g. for people at high risk, IDUs services)	65% (15)	Belgium, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Greece, Italy, Malta, Netherlands, Norway, Portugal, Romania, Slovakia
Outreach services	43% (10)	Belgium, Croatia, Denmark, Finland, Greece, Lithuania, Netherlands, Portugal, Romania, Slovakia
Routine offering in general practice	35% (8)	Belgium, Estonia, France, Malta, Norway, Portugal, Romania, Slovakia
Non-medical, non-traditional alternative settings (saunas, field visits, etc.)	30% (7)	Belgium, Denmark, Finland, France, Netherlands, Norway, Portugal
Routine offering in emergency departments	17% (4)	Belgium, Malta, Norway, Romania

2.7.4 TESSy - The European Surveillance System

In Europe, surveillance data on HIV/AIDS is collected annually from countries of the WHOE region following the guidelines published on Reporting Protocol which was a joint ECDC/WHO coordinate group since 2008. This group, along with European disease networks, ensure standardised reporting and data comparability across Member States through use of common (externally quality-assured) diagnostic and typing methods, European case definitions (in appendix A), metadata and reporting protocols [87, 105]. These data are predominantly case-based and comprise demographic, clinical, epidemiological and laboratory information.

Despite the agreed protocol, the collected data is originated in each country national surveillance systems and bears with all the heterogeneity of the national implementations. The different data collection systems take its toll on TESSy data quality.

In this work we focus on reported data from France, Italy, the Netherlands, Portugal and Poland and so it is necessary to overview their surveillance systems which is presented on Table 2.2. This sum-up focus on characteristics that, directly or indirectly may affect the

data quality of the surveillance system in terms of representativeness namely: legal coverage of the reporting; identification of the reporting entity and a summary of major historical changes in the surveillance systems. All countries described their surveillance system as case-based, comprehensive and passive [106, 103, 66]. Only in one country the reporting was voluntary whereas in the remaining countries it was mandatory. All these countries reported changes in their surveillance system. Full national coverage was achieved with delay in Italy and France started surveillance in 2003 [103].

Table 2.2: HIV-AIDS surveillance system overview [66, 106, 103]

Country	Data Source	Legal	Reported By				Comments
			Lab	Physicians	Hospital	Other	
France	FR-AIDS	Cp	N	Y	Y	N	Case-based data reported through TESSy are not exhaustive due to reporting delays and underreporting. The most recent estimates of the latter are 41% in 2007–2009 for AIDS and 30% in 2014 for HIV.
France	FR-MNOID-HIV	Cp	Y	Y	Y	N	
Italy	IT-COA-ISS	Cp	Y	N	Y	N	From 2004 until 2006, only 10 of the 22 Italian reported new HIV cases, in 2007 reported 11 regions, in 2008 reported 12 regions, 18 regions in 2009, and regions since 2012. AIDS deaths for years 2011, 2012, 2013, 2014 are not reported due to lack of updated data from the national mortality register.
Netherlands	NL-HIV/AIDS	V	N	Y	Y	N	New HIV reporting system started in 2002 including many cases diagnosed in previous years. Data prior to 2002 are from a national cohort of HIV-positive adults receiving antiretroviral therapy; 1999 data include many cases diagnosed in previous years. Control activity in 2010-2011.
Poland	PL-AIDS	Cp	Y	Y	Y	N	Control activity in 2010.
Portugal	PT-HIV/AIDS	Cp	N	Y	N	N	The database is fully case-based containing details of cases diagnosed from 1983. In 2013 and 2014, it was implemented a strategy to address underreporting and reporting delay, resulting in significant increases in the number of reported cases diagnosed between 1983 and 2012.

Y: Yes; N: No; Cp: compulsory reporting; V: voluntary

2.7.5 The Portuguese HIV/AIDS Reporting System and The Public Health System

The Portuguese HIV/ AIDS confidential case-based reporting system was performed voluntarily by physicians since the identification of the infection until 2005. Since then, it is mandatory to report all cases in any of the stages of the infection (asymptomatic, AIDS related complex and AIDS) and all the progressions including death (presented in appendix D), using case definition of WHO and the ECDC HIV / AIDS (which have minor difference but are fully compatible), identifying the probable source of transmission (Heterosexual, MSMs and IDUs) and always assuring patient's confidentiality by introducing a Soundex code. Until 2011, the forms were then sent by mail to Centro de Vigilância Epidemiológica das Doenças Transmissíveis (CVEDT), who collected, manually updated the data set, maintained, stored, analysed, produced and disseminated bi-annual reports, always keeping the data set confidential [93]. Figure 2.12 describes a simplification of the key elements of the Portuguese Surveillance System. It represents the data flow from the population under surveillance until the data recipients. This system has the general Portuguese population as target population but some key populations are more strictly looked at: HIV testing is mandatory for military and strongly recommended for pregnant women and IDUs entering

a treatment programme. Detection and primary diagnosis of infection are made by clinicians working in several settings such as: hospitals, Local Health Units, Health Centres and Grouped Health Centres. Once the case is confirmed throughout clinical and laboratory diagnosis, the form is filled in and sent. CVEDT provides data to Coordenação Nacional para a Infecção VIH - SIDA (CNSida), Portuguese working groups of the Ministry of Health, ECDC and the general public of research purpose.

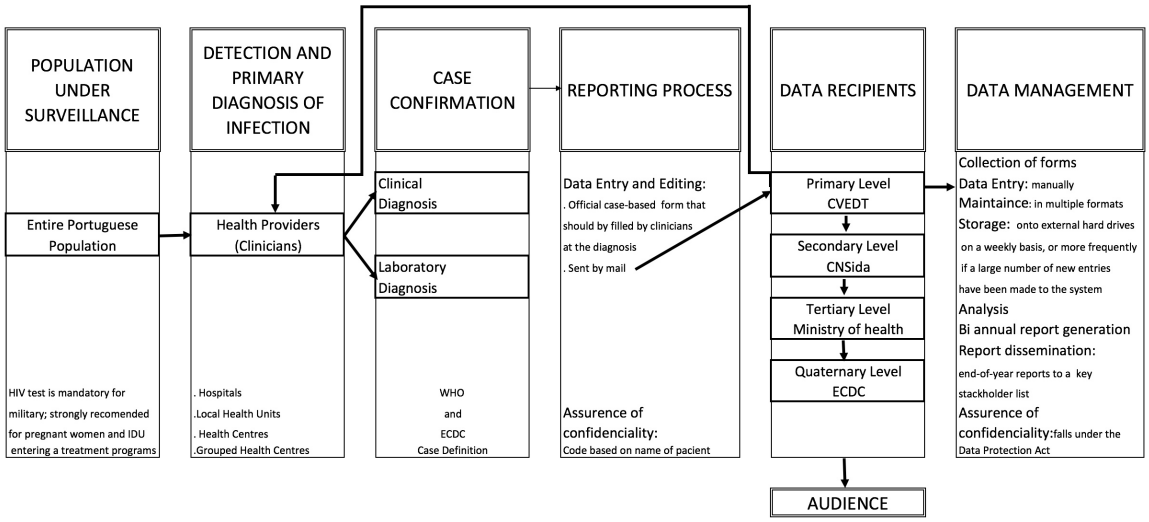


Figure 2.12: Simplification of key elements for the description of Portuguese Surveillance System based on information's reported by Mauch in [93]

In 2009, the Portuguese HIV / AIDS reporting system situation was assessed and the results were published in [93]. The study had the aim to examine the Portuguese mandatory notification system for HIV and AIDS infections as determined by the timeliness of reporting, completeness of information, acceptability by relevant stakeholders, as well as simplicity, stability and flexibility of the surveillance system.

Current national directives compel the reporting of all stages of HIV / AIDS along with its progressions and death to the Portuguese Centre for the Transmissible Diseases (CVEDT), within 48 hours after the event. Although this statutory requirement, the interviewed clinicians did not report being accountable to any time constraints. A proper notification form (appendix E) should be filled in by the patient's medical doctor. However, this procedure has only been mandatory since 2005 [107]. In 2009 Mauch study, all interviewed clinicians reported that they complete the notification form after the patient has left the office, sometimes several days or weeks later which may contribute to inaccuracies in reporting for some data variables, such as associated risk group, due to recall errors [93]. Also, according to these key stakeholders, reporting death and changes in disease status are particular problematic. After the individual paper case reports are sent to the national surveillance office, they are processed by a team of administrative staff. The form is reviewed to ascertain whether it has the minimum data variables necessary to check for a duplicate record or create a new entry. For duplicate record identification it is required at least the soundex code, date of birth and gender and for creating a new entry, additional variables are needed such as virus type, risk group, and in case of an AIDS case notification, an AIDS indicator disease from the 1993 case definition [93].

Over the years, the surveillance procedure has suffered some changes that may have altered the quality of the reports. We point out the following:

1. in 1985, the creation of HIV/AIDS surveillance system;
2. in 1988, the reporting form was altered and more variables were included;

3. in 1990, the Comissão Nacional de Luta Contra a SIDA (CNLCS) was created;
4. in 1993, tuberculosis was included as an AIDS defining disease;
5. in 1996, HAART was introduced;
6. in 2005, the notification forms were re-structured;
7. in 2009, CVEDT was re-structured (Figure 2.13)[93].

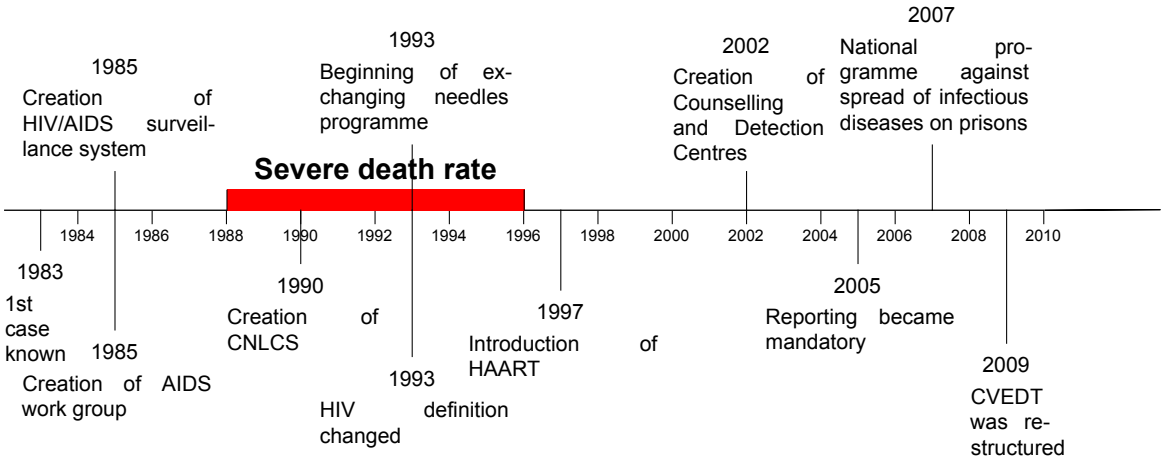


Figure 2.13: Major changes in Portuguese Surveillance System

It is also important to notice that prior to 1988, the notification form did not include important variables such as CD4 cells count, because the natural history of the infection was not fully known. And prior to 2001 the system did not record modifications to case entries, making impossible to track changes to records regarding disease progression. The entry was simply modified without tracking the date of updated diagnosis, date of updated notification or updated disease classification [93].

Similarly to other European Countries, the Portuguese Surveillance system suffers from under-reporting, under-diagnosed and reporting delay affecting timeliness and data quality. Mauch found that the average delay for all diagnosed cases was approximately 1.5 years. The under-diagnosis is less likely in late-stage infections (AIDS) as the need for medical care is much stronger and in HIV stage it depends heavily on the testing patterns of the individuals. In Portugal testing is mandatory for military and strongly recommended for pregnant women and IDUs entering treatment programs [93].

In 2013, a policy of case re-notification to improve HIV information quality was implemented, and no additional resources to assist with processing due to widespread public sector austerity in Portugal. All existing HIV positive patients in Portugal were notified irrespective of previous notification during that year resulting in tremendously increase of the number of case reports received to the national officer (from 2500 notifications per year to 23 000 notifications). The National Officer implemented a process to improve processing time but recognized that further work would be valuable to address the reporting delay between diagnosis and notification and thereby improve the overall timeliness [108]. The average

reporting delays trends display evident peaks related to this policy [103]. Two records were defined as being duplicates if it had the same soundex code. These records were verified and corrected where necessary. To prevent future duplicates a check code of EpiInfo was introduced. This function automatically searched for matching soundex codes during data entry and alerted transcribers to the potential duplicate entry of cases [108].

All data collected by the CVEDT falls under the Data Protection Act and legal action can be taken if there is misuse of the data [93].

Portuguese health care providers play an important role in care and support for HIV - AIDS patients but are also one of the principal components of the notification system; they are the primary source of health information. They work in a variety of settings of the health care system: hospitals, local health units and grouped health centres. Grouped health centres provides primary health care to the local communities, hospitals provide specialize secondary health care services and local health units' groups together the health centres and hospitals located in the same city or region in a single administrative unit, providing both primary and secondary health care.

These health providers' system is managed by the Central Administration of the Health System (ACSS) and by the five Regional Health Administration (RHA) boards implemented in mainland North, Centre, Lisbon and Tagus Valley (LVT), Alentejo and the Algarve regions 2.14.

The RHAs are responsible for the regional implementation of national health policy objectives and for coordinating all levels of health care. They work in accordance with principles and directives issued in regional plans and by the ACSS. In the autonomous regions of Azores and Madeira, health policy followed the same general constitutional principles of the National Health Service, but are implemented locally by regional governments who retained full administrative flexibility [109] (Figure 2.15).

Figure 2.14: Portuguese Health Regions

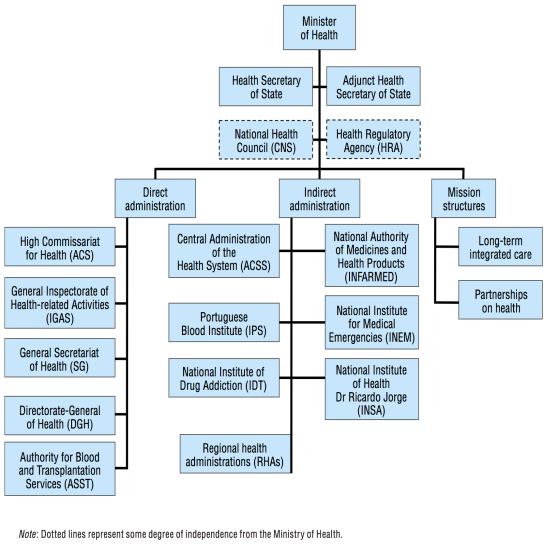


Figure 2.15: Organizational chart of the Ministry of Health from [109]

The hospital network (the number of hospitals, their location and typology) should be understood as an integrated system of health care, thought and organized in a coherent way, based on principles of rationality and efficiency [110]. The organization of the network respects national geographic diversity, equity in distribution and access to health services, different levels of intervention, primary health care, hospital care and continued care.

Portugal has adopted a "gate-keeping" system; this means that, in general, when seek-

ing for health evaluation, a patient must first go to a general practitioner (called a family doctor) and, if necessary, the primary care unit will send the patient to other levels of care, e.g. a hospital, where the specific care will be provided [111]. So, when a patient is diagnosed with HIV by a primary health care centre, usually it is referenced to see a specialist in the hospital. These special visits may also be required directly in the hospital.

Reporting HIV - AIDS cases is an administrative act that may be influenced by the factors described in the previous sections and so it is important to understand how health care providers are organized. Moreover, there are considerable inequalities in the distribution of hospital resources in Portugal [112] thus turning access and usage of health care, testing and prevention services also geographically unequal. In addition, the Health System has cost containment problems such as: accountability problems, inadequacies in the use of operational reforming tools (such as resource allocation mechanisms) and a lack of mechanisms to promote efficient behaviour [113]. Analogous results were found in primary care services [114].

Barros, in 2011, pointed out that the striking lack of nursing personnel, the scarcity of doctors in some regions and specialities (e.g. general practitioner), and the imbalance in numbers of primary care clinicians versus hospital specialists are some of the visible signs of the weakness of public health policy in the field of human resources. Moreover, the retirement in the near future of many physicians will create a shortage, as the *numerus clausus* policy applied in the past did not ensure a sufficient intake to replace them [109].

2.8 Summary

The immune system is a complex defence system that protects the host from foreign pathogenic micro-organisms. It is composed by several biological structures, such as lymphocytes T cells, and processes that react with a response to neutralize the alien and prevent the illness.

Immunodeficiency occurs when the ability of the immune system to properly respond to a threat is compromised or even absent. This state may be due to several reasons but most of the cases are acquired due to extrinsic factors such as advanced age, environmental factors and HIV infection [115].

Once the patient is infected with HIV, the immune system reacts and this process causes a short term acute illness with very few specific symptoms where, without any intervention, the virus is able to integrate into the host's cell genes. Past few weeks, the initial high virus load drastically decreases but soon starts to proliferate steadily whereas the CD4+ cells (also known as CD4+ T cell) progressively decrease over the following years. Without a strong immune system, the body becomes very susceptible to opportunistic diseases. Generally, soon after the initial HIV infection and the acute illness, the patient enters in an Asymptomatic (A) phase that may last for several years. When diseases start to appear it is said that the individual is in a Symptomatic Condition (SC) and, when severe symptoms associated to serious life threatening diseases emerge, it is said that the individual has Acquired Immunodeficiency Syndrome (AIDS).

Although HIV can affect anyone, there are key populations whose behaviours and conditions put them at greater risk of HIV exposure: injection-drug users, men who have sex with men, female sex workers and clients of sex workers [43, 44]. For the transmission of the HIV, although not sufficient, it is necessary to have exchange of certain body fluids from infected individuals. These fluids must be in contact with a mucous membrane or damaged tissue or be directly injected into the bloodstream [35, 42].

In spite of the great progresses on medication and prevention over the last years, HIV infection continues to be a major global public health issue, having claimed more than 36

million lives over the last 35 years since the recognition of the disease [39, 64]. Geographically, the burden of the disease is not equally distributed. For instance in 2015, Latvia experienced a 6.6 new AIDS diagnoses per 100,000 individuals, Portugal had a rate of 2.3 per 100,000 while the rate in the EU/EEA was of 0.8 cases per 100,000. Injecting drug use and sexual contact between men were the most reported transmission modes, so the European epidemic may be classified as concentrated in key populations.

These statistics show that while individual characteristics are necessary to the spread of the disease, they are not sufficient to explain the population level incidence. So, to reduce the burden of the infection, an endeavour and multi-factorial strategy for shaping individuals behaviours must be considered. These factors may be organized into five level hierarchical structure: individual, interpersonal, community, organizational and structural level.

Monitoring of HIV-AIDS cases, through registries, is vital to assess general health care needs and to support long-term health-policy control planning [116]. Surveillance systems have thus been established to accomplish this critical mission [117]. Typically, it depends on several processes and stakeholders, such as the population under surveillance, health care providers, laboratories and the Public Health Services, which challenges an efficient and effective detection and reporting of the diagnosed cases. This compartmentalized system may imprint several problems in the data which in turns may affect negatively the aim for which the data were collected. The most common problems are unrepresentativeness or selection bias of the population, changes in the system implementation through time, inconsistent case definitions, miss diagnoses, miss or fail to report a case, reporting delay, and errors during completion of the reporting form or data entry.

The HIV - AIDS surveillance system is a specific case of a surveillance systems due to the infection characteristics and its social repercussions. It must reflect the special transmission patterns, the long asymptomatic or with mild symptoms' latency period, the lack of affordable treatment and cure, high case fatality rates, and the social stigma associated with it [9]. More over, because of the long asymptomatic latency period and the effect of treatment, current HIV (asymptomatic) cases are hard to track and the reported AIDS cases reflect past infections that have occurred many years ago.

National surveillance systems are not identical across Europe. In addition to the common surveillance data quality issues, there are variations in data collection methods and testing policies which can introduce bias introduce bias in comparisons between countries. Such as, the comparison of HIV-AIDS case numbers between countries must be done carefully.

Portuguese HIV - AIDS confidential case-based reporting was performed voluntarily by physicians since the identification of the infection until 2005. Since then, it is mandatory to report all cases in any of the stages of the infection (asymptomatic, AIDS related complex and AIDS) and all the progressions including death, using the case definition of the World Health Organization (WHO) and the European Centre for Disease Control (ECDC) HIV - AIDS, identifying the probable source of transmission (Heterosexual, Men who have sex with Men (MSM) or Injecting Drug Users (IDUs)), always assuring patient's confidentiality by introducing a Soundex code. Until 2011, the forms were then sent by mail to Centro de Vigilância Epidemiológica das Doenças Transmissíveis (CVEDT), that collected, manually updated the data set, maintained, stored, analysed, produced and disseminated bi-annual reports, always keeping the data set confidential. Indeed, the Portuguese health care providers are one of the principal components of the notification system as they are the primary source of health information. They work in a variety of settings of the health care system: hospitals, local health units and grouped health centres but the geographical distributions of this resource is not equal.

One of the most widely recognized issue of surveillance data is the administrative reporting delay [69, 67, 66]. It can be defined as the time mediating from the identification

of the HIV-AIDS related event to its national reporting [69]. It has been mentioned that this problem depends on a number of factors such as the geographical region of the diagnosis, the calendar year, the patient's age at diagnosis and the HIV infection mode [116, 118, 119, 120, 121, 122]... Although 35 years have already passed since the identification of the disease, an accurate estimation of the reporting delay remains a challenging problem; in this thesis, we address the issue considering its multifaceted nature."

Chapter 3

Statistical Data Pre - Processing

We are living in a world of information abundance, surplus, and access. We have technologies to acquire any type of information but we still face the challenge of extracting the underlying valuable knowledge. Data analyses and mining processes may be severely impaired whenever data are corrupted by noise, ambiguity and distortions [123]. Data depends largely on its quality which is build upon the way the collection process is implemented and managed [94, 124]. Most of data analyses and mining processes focus on extracting knowledge from data; whenever the latter is of poor quality, the objective may be severely impaired and can even be beyond the scope of statistical analysis [16, 125, 126].

Data quality is an ever gradually developing concept, with roots in measurement error and survey uncertainty, that encompasses multiple disciplinary fields such as commerce, engineering, medicine, public health and policy making. This complex concept spotlights a rich set of scientific, technological and process control challenges [16, 127, 128].

There are several definitions of data quality, almost as much as the study fields. It can be defined as: *the capability of data to be used effectively, economically and rapidly to inform and evaluate decisions (with roots in decision theory field) [16]* or as *the degree to which a set of characteristics of data fulfils requirements (ISO/IEC 25012 standard, with roots in engineering and computer science) [128]* or even as *accurate, reliable, valid, timely and trusted data (with roots in integrated public health informatics network) [94].*

Despite the diversity, many definitions became aligned with the engineering, computer science and decision theory definitions. This was the situation in health information and survey research [127]. Statistics provides valuable contributions such as outliers detection, statistical data editing, probabilistic record linkage and the measurement error and survey methodology. Total quality measurement uses concepts such as multi-dimensional data quality, data quality metrics, evaluation of the user's assessment of data quality and data production maps [16].Computer science is exposed to data quality issues since organizations started collecting and storing their data electronically. So, it has well-developed technologies to address issues such as data standardization and duplicate detection and elimination, and data parsing, among others.

It is recognized that data quality is a concept with multi-dimensions (characteristics or attributes) and depends on a sustainable work-flow model integrated with minimal disruption into the day-to-day life of all the relevant stakeholders [16, 94, 128]. These characteristics are fundamental drivers that can be corrupted by specific problems with genesis on each of the steps of the work-flow. As such, the best practice process for improving and ensuring high data quality follows the so-called data quality cycle. The cycle is made up of an iterative process of analysing, cleansing and monitoring data quality [129, 130, 131].

It is very rare to find the raw data in the correct format, without errors, complete and with all the correct labels and codes that are needed for the analysis [132]. In single data collections, such as files and databases, several problems commonly arise due to mis-

spellings during data entry, missing information, lack standardization and / or validation of self-reported information, the existence of duplicate or redundant information, or the existence of anomalous events [133, 127, 94].

Data quality mining deals with detecting and removing errors and inconsistencies from data [133, 134]. Preparing and cleaning data is often more time-consuming than the statistical analysis itself [132, 125]. Moreover, developing a data cleaning procedure guided by the dimensions that characterize the quality of data is not only essential but also a quite difficult task [128]. In this chapter, it is thus presented an approach to extend, improve and systematize some state of the art methods, aiming to obtain an unified data quality methodology. The procedure applies to single file data sources without schema that may be corrupted by the most common data problems. The methodology is guided by the dimensions of data quality standards, focuses on the goal of performing reasonable posterior statistical analyses and finally formalized into an algorithm.

We will consider datasets defined according to Codd's 3rd normal form but with the constraints framed within a statistical language [125].

In section 3.1, we present an overview of the most common data quality dimensions and specific issues; in section 3.2 we provide a summary of some advanced statistical methods for data anomalies detection and data editing; in section 3.5 we preform a data quality assessment and improvement methods on the HIV-AIDS Portuguese surveillance data; the last section 3.6 contains the concluding remarks.

3.1 Data quality dimensions and issues

At a database level, data quality has a large set of dimensions but most of the researchers in informatics and official statistics have consensus on: accessibility, accuracy, completeness, consistency and timeliness [128, 135, 136]. A precise definition of each dimension can be found in ISO / IEC 25012 catalogues, in Laranjeiro *et al* or in The Quality Assurance Framework of the European Statistical System (ESS QAF) [128, 16, 136]. Briefly, we have the following:

Accuracy: The degree to which data attributes or variables correctly represent the true value of the intended object or population.

Completeness: The degree to which an entities' values correspond to all instances of the attributes.

Timeliness: Time-related dimension, reflecting if data is up to date.

Consistency: The degree to which an information object is presented in the same format, being compatible with other similar information objects or populations.

Accessibility: The degree to which data can be accessed in a specific context of use, which includes suitability of representation.

At a statistical level, we add the relevance dimension defined by the Eurostat as:

Relevance: The degree to which statistics meet the needs of current and potential research objective [136].

Data validation confronts a data set with a group of desired properties. It provides awareness for the major issues in the dataset, commonly classified as being at a structure (schema) level or at an instance-level. The latter type reflect sampling and non-sampling errors such as data processing errors (includes collection, coding and entry) or inconsistencies while the former are associated with lack of integrity constraints or poor structure design. For data sources without schema, such as single files, there are few restrictions on how data can be entered and stored, turning errors and inconsistencies more probable[133]. The most common data problems and issues as well as the corresponding affected data quality dimensions are described in table 3.1 [128, 137, 133]. We notice that if data are

inconsistent then they are also inaccurate; indeed, if values are represented in different formats and/or structures, then it is difficult to determine their true representation. So, it was extended the correspondences surveyed by Laranjeiro. Moreover, we added the dimension 'relevance' and the most commonly encountered file and statistical problems: proprietary file format, wrong variables type, tabular design, missing variables, lack of information diversity, outliers and coverage errors.

A proprietary format may contain data that is ordered and stored according to a particular encoding-scheme and so the decoding and interpretation of the stored information may only be accomplished with the use of a particular software or hardware. If the specification of the data encoding format is not released, the accessibility of the information is compromised. Also, if a variable is assigned to a wrong type, the information may be difficult to access.

Tabular design occurs when variables form both the rows and the columns and/or column' headers are values, not variables names. These questions violate the dimensions accessibility (i.e. data needs to be manually inspected and reasoning about the problem) and accuracy (i.e., the variables do not correctly represent the true population value). The issue of missing variables occurs when important variables were not included in the dataset. This compromises the dimensions completeness and accuracy, once the true object is not represented. Lack of information diversity occurs when a variable has few and under-represented unique values. This affects accessibility since the variable is not suitably represented and accuracy because the true value of the intended object may be missing. Finally, we point out the existence of outliers, corresponding to extreme values of, at least, one variable that are *far apart* from the remaining values. It may or may not be an incorrect value (although very suspicious) and by definition affects the accuracy.

Relevance may be affected by issues such as incorrect values, coverage, missing variables and lack of information diversity since the objective of the statistical methods may not be accomplished due to the lack of the necessary information. Also, illegal values, violation of logical dependencies, wrong data type or syntax violation, missing data, incorrect data, misspellings, ambiguous data, extraneous data, misfield values may impair statistical inferences and therefore affect relevance. This dimension may be affected by the presence of duplicates because they may impair representativeness of the data.

Table 3.1: Extended version of problems and data issues [128, 125]

			Data Quality Dimensions					
Problem	Description	Example	Accessibility	Accuracy	Completeness	Consistency	Timeliness	Relevance
a. Proprietary File Format	Occurs when a file is designed for a specific application / system.		X					
b. Wrong Variables Type	Occurs when a variable is assumed to be of a wrong type.		X					
c. Tabular data design	Occurs when variables form both rows and columns and/or column headers are values, not variable names		X	X				
d. Illegal values	Occurs when variables have a value that is out of range	Diagnosis date = 30/02/2006		X			X	
e. Violation of values logical dependencies	Occurs when some logical dependency between values is broken	Patient(age = 35, sex = female, HIV risk group = 'MSM')		X			X	
f. Wrong data type or syntax violation	Occurs when a value does not respect the data type constraints	Hospital = '42'	X			X	X	
g. Missing data	Occurs when data is not present or coded with a dummy value	Probable year of infection = 9999, 'Not mentioned'		X	X		X	

(Table continues...)

Table 3.1: Extended version of problems and data issues [128, 125]

Problem	Description	Example	Data Quality Dimensions: Accessibility	Accuracy	Completeness	Consistency	Timeliness	Relevance
h. Incorrect data	Occurs when data contain valid values that do not correspond to real values	HIV risk group = 'Heteressexual' instead of 'IDUs'		X			X	X
i. Misspelled	Occurs when data are misspelled	Nationality = 'Portgal' instead of 'Portugal'		X			X	
j. Ambiguous data	Occurs when data values can have different interpretations. It can be due to abbreviations or an incomplete context	Hospital = 'Maria', it can be 'Hospital Santa Maria' or 'Hospital Santa Maria Maior' or even 'Hospital Santa Maria Pia' (Hospitals in Portugal)	X	X			X	
k. Extraneous data	Occurs when additional data is represented	Name = 'Mrs. Maria'	X	X		X	X	
l. Outdated temporal data	May be valid for a time point or interval	Hospital = 'Desterro' (this hospital was closed in 2007)		X			X	
m. Misfielded values	Occurs when the data values are stored in the wrong column	CD4 count in the variable of T4/T8 ratio	X	X	X	X		X
n. Duplicates	Occurs when the same data appear more than once but with a different identifier and may have contradicting information	Patient1(ID = 'AERER', age = 31, sex = 'male', risk group = 'MSM') and Patient2(ID= 'AERER', age = 31, sex = 'male', risk group = 'IDU')	X	X			X	
o. Outliers	Occurs when a variable has an outlier. It may or may not be an incorrect or dummy value (although they are very suspicious)	Age=88,age=99		X				
p. Missing Variables	Occurs when an important variable is missing.	Individual records of HIV patients without clinical variables.		X	X			X
q. Lack of information diversity	Occurs when a variable has few and under-represented unique values.	All observations of variable HIV type equal to HIV1.	X	X				X
r. Coverage errors	Occurs when a unit in the sample is incorrectly excluded or included, or is duplicated in the sample.			X	X*			X

* Just in case of excluding a sample.

It is important to point that data validation adds value to the dataset by rising awareness about the presence of issues that may impair its usefulness for the intended purposes [138].

3.2 Procedure for improving data quality

Data Quality Mining (DQM) can be defined as the deliberate application of data mining techniques for the purpose of data quality improvement [134]. The main goals in any data quality improvement process are: the detection, the explanation of the source and correction of deficiencies such that the 'improved' dataset is close as possible to the original collected data. Other important goal is to confidently allow for posterior statistical analysis.

In this chapter we follow Wickham and Jogen definitions of dataset [125, 132]: it is a collection of values (qualitative or quantitative), each belonging simultaneously to a variable and an observation. A variable contains all values that measure the same underlying attribute across units. An observation contains all values measured on the same unit across attributes. This is Codd's 3rd normal form but with the constraints stated within a statistical framework [125].

The central problem for data improvement is how to make data consistent, accurate, accessible, timeliness, complete and relevant while keeping the edits to as few as possible,

respecting the constraints and context [132, 139, 140].

A common approach to address the problem of dirty data is to apply a set of data quality rules or constraints over a target database, to 'detect' and to eventually 'repair' data issues [139], but first one must be able to read and understand the data. After that, four stages can be identified: data structuring, data validation, error localization and repair.

3.2.1 Data Validation

The credibility of the data may be assessed by data validation activities whose negative outcome guarantees the data have quality issues, mainly accuracy and consistency dimensions. Data validation can be defined as follows:

Definition 1 *Data validation is an activity verifying whether or not a combination of values is a member of a set of acceptable combinations [138, 141] .*

The set of 'acceptable combinations' may be a set of possible values for a single variable or a set of valid combinations of values for an observation, variable, or larger collection of data [138]. It is an iterative procedure based on the tuning of validation or edit rules that will converge to a minimal set that must be necessarily satisfied [141]. Moreover this set must be complete, concise and consistent [142].

Completeness may be defined as

Definition 2 *Completeness is the extent to which prior knowledge about the data set has been expressed in terms of a set of validation rules or the extent the variables in a data set are covered by the rules in a set.*

The prior knowledge includes rules that arise from physical or logic facts and from domain experts. When the influence on the data values are due to facts or events that can only be stated imprecisely it is difficult to state it on validation rules.

This property is hard to quantify but may be assessed by knowledgeable peers review or by checking whether at least one of the explicitly defined rules is valid (assuming that the set of rules is on a standard and minimal form) [142].

A concise set of validation rules can be defined as

Definition 3 *A concise set of validation rules is a irreducible set that contains the minimal set of rules that must be satisfied.*

A set of validation rules divides the space of all possible records into a valid, or acceptance region, and an unacceptable region. A rule is redundant in the set if removing it from the set does not alter the acceptance region [142]. Linear programming is a common approach for redundancy removal.

Finally, a consistent set of validation rules is a set with a non-empty acceptance region. Inconsistency occurs when a rule set contains a rule that is contradictory in itself [142]. Common strategies for finding inconsistencies in a set of rules are the Fourier-Motzkin elimination variables for rules involving numerical variables, and multivalent resolution for rules involving categorical variables.

Statistical data editing is the automated process of stepping through the data observations and correcting them whenever they violate the validation rules [16]. The validation activities may be grouped, from a research perspective, in six ordered levels starting with simple single value checks and moves to more complex checking involving more observations, variables and data sets [143, 144]:

Level 0: Validation of the Information Technology structural requirements. At this level the data set is checked format and file structure and variables types (issues **a,b** and **c** from Table 3.1).

Level 1: Validation for consistency of single data points. For example comparing a single data point with constants. At this level the data set is confronted with rules for checking the presence of issues as described in **d, f, g, i, j, k** and **m** on Table 3.1.

Level 2: Validation for consistency of multiple variables of the same statistical object. At this level the data set is confronted with rules for checking the presence of issues as in **e** above.

Level 3: Validation for consistency of multiple statistical objects of the same variable. At this level the data set is confronted with rules for checking the presence of issues such as **o** and **q**.

Level 4: Validation for consistency of repeated measures of the same variable and same statistical object. At this level the data set is confronted with rules for checking the presence of issues such as **l**.

Level 5: Validation for consistency of multiple variables and multiple statistical objects. At this level the data set is confronted with rules for checking the presence of issues such as **n**.

Level 6: Validation for consistency based on the comparison of the file content with the content of other files. By other files we mean other versions of exactly the same file, same data set but referring to a different time period, different data set but collected from the same domains or even from correlated domains. At this level the data set may be confronted with rules for checking the presence of issues such as **h, p** and **r**.

These proposed steps for the validation process are represented in Figure 3.1.

In these chapter we followed the definition of validation rules given by Jonge *et all* in [145].

Denote by $v_i\,(i = 1, \ldots, m)$ a categorical variable and by $x_j\,(j = 1, \ldots, n)$ a numerical variable. The finite set of observed categories of the i^{th} categorical variable is represented by D_i, where

$$D_i = 1, 2, \ldots, n_i.$$

In a general form, a statistical object r may be written in the form

$$r = (v_1, \ldots, v_m, x_1, \ldots, x_n) = (\mathbf{v}, \mathbf{x}) \tag{3.1}$$

taking values in $D_1 \times D_2 \times \ldots \times D_m \times \mathbb{R}^n = D \times \mathbb{R}^n$ [145].

Usually rules for categorical variables are defined negatively by stating a subregion of D with the disallowed values, while linear validations define allowed regions in $\mathbb{R}^n$. It is possible to use a complementary approach in which rules are expressed in "negative form", then validation is done by verifying that a pre-defined non - acceptable combination of values do not occur [141]. So, for a single edit the region G with rules in a negative form can be defined as:

$$G = \{r \in D \times \mathbb{R}^n : \mathbf{v} \in F \wedge \mathbf{x} \in P\} \tag{3.2}$$

with $F \subset D$ representing the values that are considered as invalid. Note that $F = F_1 \times F_2 \times \ldots \times F_m$ such that $F_i \subset D_i$. And P is a convex subset of $\mathbb{R}^n$ defined by a set of k linear restrictions $(k = 1, \ldots, n)$ of the form $A\mathbf{x} > b$.

The positive reformulation of 3.2 is:

$$\begin{gathered} \neg\,(v \in F \wedge x \in P) \\ \neg\left(v \in F \wedge a_1^T x_1 > b_1 \wedge a_2^T x_2 > b_2 \wedge \ldots \wedge a_n^T x_n > b_n\right) \\ v \in \overline{F} \vee a_1^T x_1 \leq b_1 \vee a_2^T x_2 \leq b_2 \vee \ldots \vee a_n^T x_n \leq b_n \end{gathered} \tag{3.3}$$

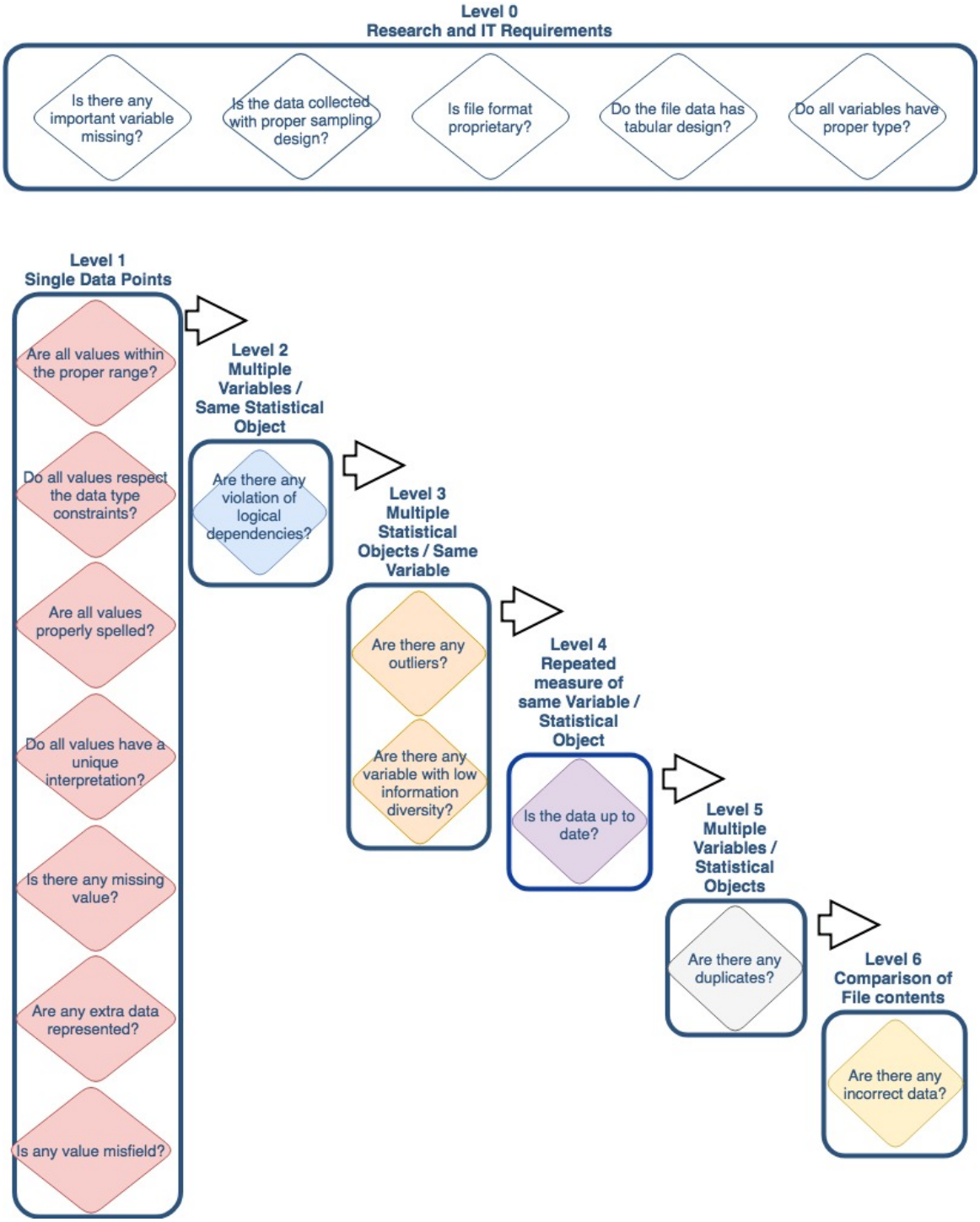


Figure 3.1: Validation Process

where $a_i \in \mathbb{R}, (i = 1, \ldots, n)$.

In this formulation, it is said that r is valid if and only if

$$r \in \overline{G} \Leftrightarrow \vee_{i=1}^{m} v_i \in \overline{F_i} \vee \vee_{j=1}^{k} a_j^T x \leq b_j$$

As Jonge pointed out in [145], the rule defined in equation 3.3 can be formulated as a 'conditional validation' by using the implication replacement rule from propositional logic which states that $\neg p \vee p$ may be replaced by $p \Rightarrow q$. If we limit equation 3.3 to a single condition,

$$\mathbf{v} \in F \Rightarrow \mathbf{a}^T \mathbf{x} \leq b \tag{3.4}$$

Validation checking, then, involves evaluating $\mathbf{v} \in F \Rightarrow \mathbf{a}^T \mathbf{x} \leq b$ and comparing the result with the valid-values set $\bar{G}$. If the rule cannot be evaluated then the checking returns NA. So validation checking is a surjective function of a generic statistical object r:

$$V : D \times \mathbb{R}^n \rightarrow \{0, 1, NA\} \tag{3.5}$$

where 0 is interpreted as invalid, 1 as valid and NA when the rule cannot be evaluated. The valid region of $D \times \mathbb{R}^n$ is defined as the preimage

$$\bar{G} = V^{-1}(1) = \{r \in D \times \mathbb{R}^n : V(r) = 1\} \tag{3.6}$$

3.2.2 Error Localization

Data editing is a complex process that often involves cross-variable and interrelated contingency rules. Fellegi and Holt formulated a theoretical model for this process with the following goals [146]:

1. in each record or observation, data should satisfy all validation rules by changing the fewest possible items (variables or fields);
2. imputation rules should be derived automatically from validation rules;
3. when imputation is necessary, it is desirable to maintain the marginal and joint frequency distributions of variables.

The problem of finding the variable(s) violating the validation rules and needs to be corrected can be stated as following:

Problem 1 *Given a statistical object r with n variables, subject to a number of possibly multivariate validation or consistency rules, find the smallest (weighted) subset of variables, such that after replacement of their values, r violates no rules [145].*

De Waal in [147] presents the mathematical formulation of the above problem as follows:

$$\text{Minimize } \sum_{i=1}^{m} w_i I_i + \sum_{i=1}^{n} w_{m+i} I_{m+i}; \tag{3.7}$$

$$\text{subject to } \mathbf{v} \in F \Rightarrow \mathbf{a}^T \mathbf{x} \leq \mathbf{b}. \tag{3.8}$$

where $I_i = \begin{cases} 1 & \text{if field } i \text{ is to be changed} \\ 0 & \text{otherwise} \end{cases}$ and w_i is a reliability of variable i given by field experts. A small value of w_i indicates that the variable is more likely to have an error. So, this is an optimization problem over $m + n$ binary decision variables that indicate which variables in a record should be changed [145]. As Jonge pointed, this problem is NP-complete where "the search space related to the minimization problem grows exponentially with the number of fields". Moreover, the problem is complicated by the existence of implicit rules (rules that are logically or arithmetically derived from explicit rules).

There are several algorithms for solving this problem, the two most popular being the branch-and-bound and mixed-integer programming, described in [148].

3.2.3 Repair and Improve

In this subsection, methods for repairing the main issues presented on Table 3.1, guided by the validation steps given on page 40, will be over-viewed. Note that the correction steps are not in the exactly same order as the checking steps; more precisely, all the imputation methods are done after all the errors have been addressed.

It is recommended that data is in a text-based format because this has several favourable properties over other formats, including being human readable, the possibility to represent any type of value, and avoiding underlying stored data issues.

More than 80% of all data is unstructured [124]. For statistical analysis it is desirable that the values are organized in a rectangular table with rows and columns such that:

1. each column corresponds to a single variable, with a proper label,
2. each row has just one observed case, and
3. each type of observational unit forms a table.

A bad table design can be due to: column headers representing values and not variable names, one column containing multiple variables, variables being stored in both rows and columns, same table storing different observational units or single observational unit being stored in multiple tables [125].

These issues are out of the scope of the current work but often require manual inspection of the data(s) file(s), reasoning about the problem representation on the dataset and the identification of what needs to be re-structured keeping the focus on the aim of the analysis and domain expertise.

Once the file is in a type agnostic text format, like csv or tsv, and after being properly structured, variables' parsing has to be performed; this means converting each variable from the text string representation to memory typed values, ready to be manipulated with statistical tools like R or SPSS.

Since data improvement is a resource consuming process, it is important to chose wisely which variables need to be cleaned taking into consideration the context relevance. For the selected variables we should proceed with a uniformization of the values. A simple scan for unique values and its corresponding distribution in data can detect several problems such as: misspellings, extraneous data, misfield values, wrong data type or syntax violation, ambiguous data and illegal values.

The first two issues may be addressed by simply editing the value to the corrected form. The misfield values should be placed in the correct variable leaving the original cell empty. In the case of syntax error, one must follow the same procedure as that of misspelling errors, but if the error is of wrong data type that data editing is more difficult. The first solving approach should be consulting the metadata for guidance and consult the data domain experts. If this does not solve the problem, then one must find support among other variables and look for logical dependencies with other variables (this issue will be addressed later in this chapter). If this still does not prove to be helpful then one should simple delete the value. A wrong value is worse than no value at all because it may introduce unwanted bias. The procedures for ambiguous and illegal values should be the same as those for the wrong data type.

An observation (or set of observations) which appears to be inconsistent with that set of data is called an outlier [132]. For detection of outliers, several approaches have been developed, such as error bounds, tolerance limits and control charts; model-based regression depth and residual analysis, distributional representations and time-series analysis.

Outliers can appear from several different mechanisms or causes: they can be simply an error or arise from the inherent variability of the data. When an observation is clearly an error, simply delete the value but, if it is legitimate or of unclear source, then the problem

becomes difficult. One way for keeping an outlier is to transform the variable either by applying a function or by truncation to the closest extreme. If none of these transformations turn out to be helpful, then one may need to use “robust” parameters estimate methods for data mining[149].

Violations of logical dependencies between variables are easily detected with data visualization tools and with cross tabulations. Each dependence between variables, also known as editing rule, require the definition of a proper pre-specification of specify domain-knowledge-based constraints.

As a violation may be resolved in more than one way, an immediate question is which one to choose? Solutions may include repairing values that require the least number of operations, or repairing values according to a pre-specified cost model. Many cost models have been proposed but they rely on the same idea: to associate weights according to the user’s confidence on the accuracy of the values [140] . For ”expensive“ editing, the rule is to delete the case.

After the cleaning process, one should turn his/her attention to missing values. Since attempts to recover missing values may impair inferences (the main goal of a statistical procedure), a missing value treatment cannot be properly evaluated apart from the modelling, estimation, or testing procedure in which it is embedded [150]. Several methods for data imputation have been proposed and they can be divided into two categories: single variable or considering relationships among variables [126]. For single variable methods it is assumed that the distribution of the missing values is the same as the non-missing values and so the missing values are replaced by the non-missing values mean, median or other point estimates. The methods that consider relationships among variables use regression (parametric) or propensity scores (nonparametric). These techniques assume that the explanatory variables are non-missing. If the proportion of observations having any variable missing is less than or equal to 5% then missing values imputation does not have high impact and a single imputation method can be used [151]. Otherwise, when the rate of missing values is high, the chosen imputation method ’will exert a high degree of influence over the results’ [150]. To guide the choice of the imputation method the researcher must firstly identify the pattern of missing values and use a method that considers the relations between the variables. After any imputation procedure it is necessary to perform a sensitivity analysis, comparing the pre-imputation with the post-imputation scenarios.

Methods for adjusting / repair outdated temporal data will be discussed in Chapter 4.

Duplicated records may be introduced in a data set through several different mechanisms. Typically, by typographical and data entry errors. An individual’s information may change over time with some life events like moving or deliberately reporting false informations [152]. Due to these idiosyncrasies, the previous data cleaning and standardization steps are critical.

The duplicate detection is typically performed by applying similarity functions to pairs of observations. If the values of two records are sufficiently similar, they are assumed to be duplicates’ [153]. This task starts with the definition of a search string key composed by, at least, one variable. Records that share a key are, without a doubt, duplicates. But some times, the variables in the genesis of the key are corrupted by different formatting conventions, abbreviations, typos ... In that case it is recommended to use algorithms of approximate string comparison. These algorithms allow for character deletion, insertion, substitutions and transpositions, common sources of differences when comparing strings. The identification of these suspected duplicate records is a difficult task but once they are flagged they can be processed manually or in a semi-automated manner [154, 16].

Some common and successfully similarity measures that handle typographical errors well are: edit-based distances (Hamming, generalized Levenshtein, the longest common sub-string, optimal string alignment, and generalized Damerau - Levenshtein), q-gram based

distances (q-gram, Jaccard, and cosine) and the heuristic Jaro and Jaro-Winkler distances. The edit-based distances are determined by counting the number of fundamental operations necessary to transform one string to another. Operations may include substitution, deletion, insertion of a character or transposition of characters. When more than a single operation is needed, it may be useful to assign weights to the different operations, 'for example to make a transposition contribute less to the distance than character substitution' [155]. 'Distances based on q-grams are obtained by comparing the occurrence of q-character sequences between strings. Heuristic measures have no strong mathematical underpinning but have been developed as a practical tool with a particular application in mind' [155]. One example is the Jaro distance that was originally developed for the purpose of linking records based on inaccurate text fields, by the U.S. Bureau of the Census, and it is very efficient on statistical matching problems involving relatively short string [155].

Methods for detecting duplicates consisting of multiple variables can be divided into two categories:

1. Approaches that rely on training data to "learn" how to match the records; these include probabilistic approaches and supervised or semi-supervised machine learning techniques.
2. Approaches that rely on domain knowledge or on generic distance metrics to match records; these include deterministic linkage methods using distance metrics appropriate for the duplicate detection task [154].

Successful probabilistic approaches in duplicate detection comprise technique such as Bayes decision rule for minimum error, for minimum cost or with reject region. Supervised or semi-supervised methods include: Classification and Regression Trees (CART) algorithms, Support Vector Machines (SVM), Supervised Clustering (SC) techniques, Semi-Supervised Probabilistic Relational Models (SSPRM), Markov Chain Monte Carlo (MCMC), Sampling Algorithm (SA) among many others. Approaches using distance metrics include the Hungarian Algorithm (HA) Hungarian Algorithm, Successive Shortest Paths algorithm, experts rule-based bootstrapping, techniques based on clustering, hierarchical or graphical models for learning to match record pairs, just to cite a few [154, 156].

This task is highly data-dependent and therefore choosing a detection technique is similar to model selection and performance prediction for data mining tasks [154, 156]. However, it has been demonstrated that deterministic approaches have high validity and reliability and has been employed successfully in multiple updates of the SEER (Surveillance, Epidemiology and End Results)-Medicare linked dataset. "The algorithm consists of a sequence of deterministic matches using different match criteria in each successive round" [152].

If the records detected as suspicious duplicates match in every variable, then just delete one of those records. But if duplicated records have contradicting information the task is much more difficult requiring a manual clerical review to decide their final match status [156]. If they are to be considered as duplicates the next task is to decide what to do with them: delete both, one or merge the records. If the decision is to keep just one, the natural question is which one. The decision of the record that must be deleted can be based on a cost model ranking the records from the most to the least reliable. This cost model can take into consideration the number of data quality issues contained in each record and a domain expert weight about the reliability of each variable.

For correcting ambiguous data and missing variables it is useful to check different versions of the same data set, or to match the data with correlated data sets from the same or different sources or even different domains.

Incorrect data rising from random intentional errors may be treated as outliers but if these errors are systematic, it is very hard to detect and correct them.

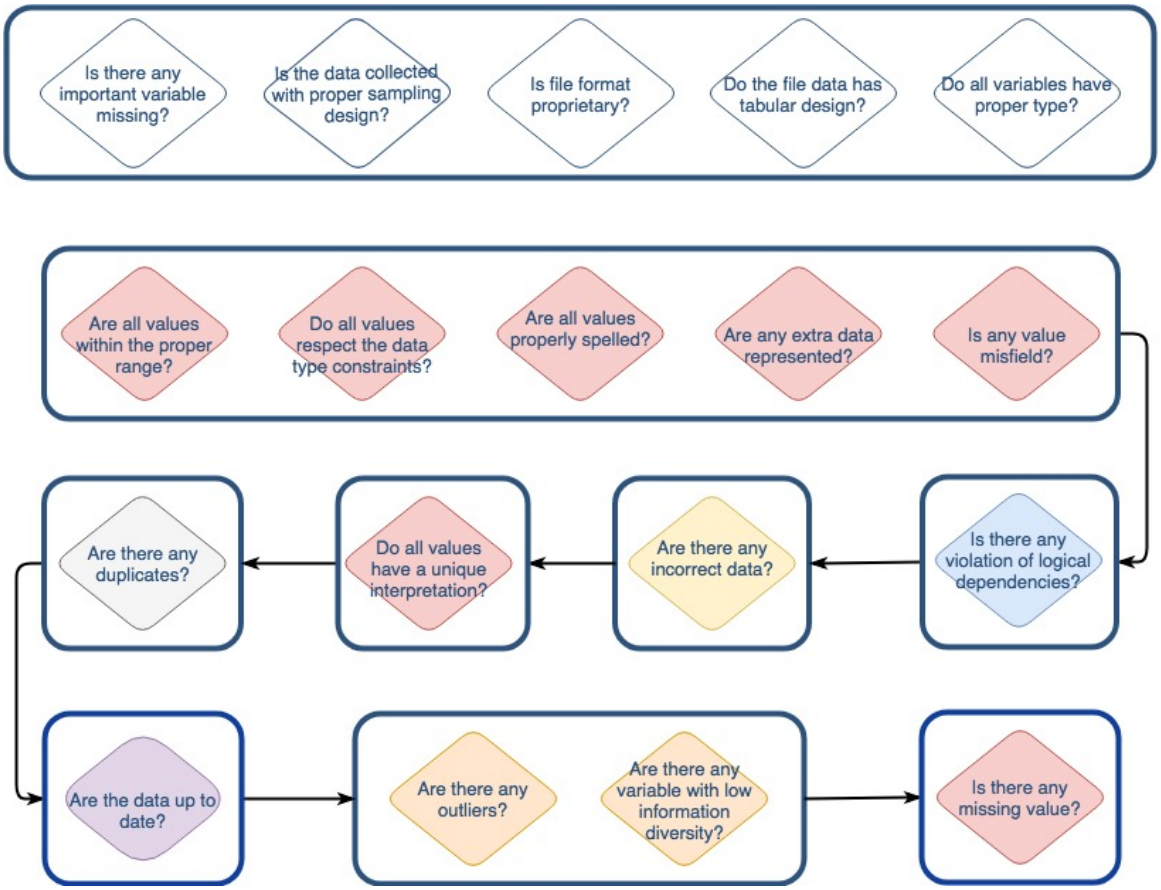


Figure 3.2: Repair and Improve Process

Coverage errors can be minimized by capture - re-captures techniques, matching the data set with a correlated data set from a different domain.

A relevant audit is if any important variable is missing in the dataset. If so, the main goal for which the data was collected may be irreparably compromised.

In this section, it is proposed a macro sequence across the levels of validation, which follows from level 0 until 6 but the contained errors checking may be of any order. The repair process, however, must follow a more strict order, the last addressed errors being the ones for which the correction involves imputation and outliers handling due to the obvious influence over the results of a subsequent statistical analyses. The steps proposed for data repairmen and improvement are presented in are present in Figure 3.2.

Once the repair and improve process is ended, data is then ready for statistical analysis and model selection. The above procedure is formalized up in algorithm 3.1 below.

3.3 Surveillance Data

ECDC and WHO jointly collect HIV - AIDS surveillance data covering the full WHO European Regions, i.e. the 31 European Union / European Economic Area (EU / EEA) countries and 23 countries outside the EU / EEA, through TESSy - a web-based data submission, data storage and dissemination platform. This data collection is guided by a Reporting Protocol designed to improve user-friendliness by introducing a uniform structure, standard variables, data formats and codes [105].

The stored anonymized data may be accessed for research purposes / tasks in the public interest upon request which should be placed to the competent national authority.

Algorithm 3.1 Data quality improvement

procedure Issue detection and repair
Input: A dataset A
Output: An improved dataset
alpha ← Data Explore(A) {
if A satisfies all necessary specifications **then**
alpha =`True`
else
alpha =`False`
end if}
if alpha=`True` **then**
B ← Data Structure(){
Ensure: Each variable in a column, each observation in a row, each observational unit in a table.
B = arranged A }
C ← Data Parsing(B){
for all Variables in B **do**
Convert variable type to the proper one
end for
Store variable in C }
D ← Select Important Variables(C){
Select the variable to be clean according to the context
Store variable in D }
if D not empty **then**
E ← Uniformization(D){
for all Observations of text variables **do**
if value is affected by an issue of type d., f., i., k. or m. **then**
Edit value to correct form
end if
Store values in E
end for}
F← Missing Values Imputation(E){
for all Variables in E **do**
Apply techniques for missing values imputation
Perform sensitivity of the results to departures from assumptions analysis
end for
Store results in F }
G ← Identification and Correction of Outliers(F){
for all Numerical variables in F **do**
Apply techniques of outlier detection
Apply techniques to minimize the effect
end for
Store the results in G }
H ← Validate Logical Relationships(G){
Apply validation rules to G
Determine the minimal set of variables to be changed
Correct any violation
Store the results in H }
I ← Missing Values Correction(H){
for all Variables in H **do**
Determine missing values pattern
Impute missing values
end for
Store the results in I }
J ← Outdated Data Correction(I){
for all Variables in I **do**
Determine if data is outdated
Correct any violation
end for
Store the results in J }
K ← Duplicate Detection and Correction(J){
Apply similarity measures between observation
Apply algorithms to find matches in datasets
Correct the duplicates
Store results in K }
L ← Ambiguous, Incorrect or Missing Variables(K){
Determine if there exists missing variables, incorrect or ambiguous data
Match the file with correlated information
Store the data in L }
M ← Data Transformation(L){
Compute important variables
Store variables in M }
else
Collect new data
end if
else
Collect new data
end if
end procedure

In the particular case of the Portuguese surveillance system, the data is collected through a standardized notification form for HIV infections that was revised several times since the identification of the disease. The most recent version, currently in use, was introduced in 2005 and can be found in appendix E). “A single notification form is used for reporting both HIV and AIDS infections, as well as any change in disease status, including death, for both adult and paediatric cases. The notification form consists of 13 categories, with approximately 100 potential data field” [93]. Also, data is accessed upon request for research purposes / tasks in the public interest.

3.4 ECDC dataset structure: HIV / AIDS case-based record type

The dataset for case-based combined HIV / AIDS surveillance consists of a set of 33 variables, 17 of which are mandatory and the remaining 16, optional. The mandatory variables represent a minimum standard designed to provide an overview of the ECDC epidemic and to improve data completeness and accuracy across Europe”. The optional variables are aimed at countries that are able to complete them and are aspirational to other countries for using this as guide to design and enhance national surveillance systems [105].

To accomplish the main objectives of this research, and under the ECDC guidance, we restricted our attention to the variables presented in Table 3.2. Variables’ values and used validation rules are described in Table 3.3. With this respect, we have used the following notation:

- Errors - indicates that the structure of the file should be corrected. The file is rejected, should be reformatted and then uploaded again;
- Warning- is always combined with a course of action to improve data quality;
- Remark - informs the user that steps should be taken to improve data quality [3].

Table 3.2: Overview of the available revised set of variables for case-based HIV-AIDS surveillance. Source [105]

Variable	Repor Type	Description
TESSy System Related Variables		
1. RecordType	M	Structure and format of the data (case based reporting and aggregate reporting). It specifies what data values TESSy expects to receive.
2. RecordTypeVersion	O	The version of the record type defines the current structure of the data reported. If the original dataset for any particular disease is changed, the versioning will change according to increasing numbering.
3. Subject	M	The subject describes the disease associated to the reported case(s).
4. DataSource	M	The data source (surveillance system) from where the record originated.
5. ReportingCountry	M	The country reporting the case.
6. DateUsedForStatistics	M	This is the date used by the national surveillance institute or organisation in the national HIV/AIDS case reports and other official statistics.
Diagnosis Information		
7. DateOfDiagnosis	M	The date of first HIV diagnosis; clinical or laboratory diagnosis.
8. DateOfNotification	M	This is the date on which the HIV case was notified for the first time to the the surveillance system in the reporting country.
9. HIVType	M	Describes the type of HIV infection
10. HIVStatus	O	Information on previous positive test results, prior to the first time of reporting. This variable allows cases that are “newly diagnosed” to be distinguished from cases who had a positive HIV test in the past but are tested and/or reported for the first time in the country (i.e. transfer of care).
11.Stage	O	This variable specifies the clinical stage at the time of HIV diagnosis.
12. Transmission	M	Most probable route of HIV transmission.

M: Mandatory; O: Optional (Table continues...)

Table 3.2: Overview of the available revised set of variables for case-based HIV-AIDS surveillance. Source [105] - continue

Variable	Repor Type	Description
13. FirstCD4Date	M	Date of first available CD4 cell count.
14. AcuteInfection	O	Evidence of recent infection, aside from the recent infection assay result. An infection can be considered to be recent if a patient presents with seroconversion illness, has a negative HIV test within 6 months of diagnosis or has evidence from p24 antigen or Western Blot assays.
Demographics		
15. Age	M	This is the age at diagnosis of the person in years as reported in the national system of the Member State.
16. Gender	M	Gender of the infected person.
17. RegionOfOrigin	O	Region from which the case originates.
Clinical Information		
18. ART	O	Was the patient receiving antiretroviral therapy at the date last seen for care?
19. CD4cells	O	CD4 cell count at time of diagnosis.
20. DateOfAIDSDiagnosis	O	The date of first AIDS diagnosis; clinical or laboratory diagnosis.
21. AIDSIndicatorDisease	O	AIDS indicator disease at the time of AIDS diagnosis occurring within two consecutive months from the date of AIDS diagnosis.
Death		
22. Outcome	O	The vital status of the patient: 'Alive' or 'Died'.
23. DateOfDeath	M	Date of death due to any cause.
24. DeathCause	M	Information on whether the case is alive or deceased (due to AIDS-related and non-AIDS related causes).

Table 3.3: Values and validation rules of the revised set of variables for case-based HIV-AIDS surveillance. Source [105]

Variable	Codes, Allowed Formats and Validation Rules
TESSy System Related Variables	
1.RecordType	HIV
2.RecordType-Version	
3.Subject	HIVAIDS = HIV diagnoses case-based, including AIDS.
4.DataSource	FR-HIVAIDS = France Surveillance; IT-COA-ISS = Italy Surveillance; PT-HIVAIDS = Portugal Surveillance NL-HIV/AIDS = Netherlands Surveillance; PL-HIV = Poland Surveillance
5.ReportingCountry	FR = France; IT = Italy; NL = Netherlands; PL = Poland; PT = Portugal
6.DateUsedStatistics	yyyy, yyyy-Qq, yyyy-mm, yyyy-ww, yyyy-mm-dd
Diagnosis Information	
7. Date of Diagnosis	yyyy, yyyy-Qq, yyyy-mm, yyyy-ww, yyyy-mm-dd (Error) Date of diagnosis must be after 1970.
8.DateOfNotification	yyyy, yyyy-Qq, yyyy-mm, yyyy-ww, yyyy-mm-dd
9.HIV Type	HIV1 = HIV1 only; HIV2 = HIV2 only; HIV12 = HIV 1 and HIV 2 (co-infection); Unk =Unknown
10.HIVStatus	PREVPOS = Previous HIV positive; NEG = Negative; UNK = Unknown
11.Stage	ACUTE =Acute HIV infection; AIDS = AIDS; SYMPNONAIDS = Symptomatic non - AIDS; ASYMP NONAIDS = Asymptomatic Non - AIDS, not further specified; UNK = Unknown (Remark) In each patient register, Stage should be reported. (Warning) If Stage is reported as acute infection, then DateOfDiagnosis should be reported. (Warning) If Stage is reported as AIDS, then DateOfAIDSDiagnosis should be reported. (Warning) If Stage is not reported as AIDS, then DateOfDiagnosis should not be later than DateOfAIDSDiagnosis
12.Transmission	HETERO = Heterossexual contact; TRANSFU = Transfusion recipient; MSM = MSM/homo or bisexual male; HAEMO = Haemophiliac patient; MTCT = Mother-to-child-transmission; IDU = Ever injected drugs; UNK = Unknown or undetermined including case of occupational exposure (Error) If transmission category is men how have sex with men, then gender shouldn't be female (Remark) If transmission category is reported as transfusion of blood or its component due to haemophilia, then gender is not very often female (Warning) If the person is younger than 13 years old, transmission is often mother to child (Error) If transmission category is reported as MTCT, stage cannot be acute HIV infection
13.FirstCD4Date	yyyy, yyyy-Qq, yyyy-mm, yyyy-Www, yyyy-mm-dd, UNK, NA
14.AcuteInfection	EV24ANT = Evidence from p24 antigen; EVWBLOT = Evidence from Western Blot; NEGTEST=Last negative test within 6 months of HIV diagnosis; SEROILL = Seroconversion illness; NA = Not applicable (not acute infection); UNK = Unknown
Demographics	
15.Age	(Warning) If the person is younger than 13 years old, transmission is often mother to child (Remark) It is unlikely that the age is more than 80 (Age <=80).
16.Gender	F = Female; M = Male; O = Other (e.g., transsexual); Unk = Unknown (Error) If transmission category is men who have sex with men, then gender should not be female. (Remark) If transmission category is reported as transfusion of blood or its component due to haemophilia, then gender is not very often female

(Table continues...)

Table 3.3: Overview of the revised set of variables for case-based HIV-AIDS surveillance. Source [105] - continue

Variable	Codes, Allowed Formats and Validation Rules
17.RegionOfOrigin	ABROAD = Abroad, sub continent unknown; AUSTNZ =Australia and New Zealand; CAR=Caribbean; CENTEUR = Central Europe; EASTASIAPAC = East Asia and Pacific; EASTEUR = East Europe; LATAM=Latin America; NORTHAFRMIDEAST=North Africa and Middle East; WESTEUR=West Europe; EUROPE = If not in West, central or Eastern Europe, should be reported in Europe; NORTHAM = North America; REPCOUNTRY = Same as reported country; SOUTHASIA=South East Asia; SUBAFR = Sub Sahara Africa; NORTHAFRMIDEAST = North Africa and Middle East; Unk=Unknown
Clinical Information	
18.ART	Y = Yes; N = No; UNK = Unknown (Remark) If ART is Y, last attendance date should be reported.
19.CD4cells	A numerical values of the CD4 from 0 to 6000 or unknown (UNK) (Remark) Usually CD4 cell count varies in a range from 0 to 1500 and due to rare extremely values upper high limit is set to 6000 per cubic millimetre.
20.DateOfAIDSDiagnosis	yyyy, yyyy-Qq, yyyy-mm, yyyy-ww, yyyy-mm-dd (Error) Date of AIDS diagnosis must be after 1970. (Remark) If death due to AIDS is reported, patient should also have a DateOfAIDSDiagnosis.
21. AIDSIndicator Disease	31 codes of the AIDS diseases defined in Appendix C (Warning) If AIDS indicator disease is reported, DateOfAIDSDiagnosis should be reported.
Death	
22.Outcome	A = Alive; D = Died; UNK = Unknown (Warning) If it is known that a person died, it is usually expected that the Date Of Death is reported.
23.DateOfDeath	yyyy, yyyy-Qq, yyyy-mm, yyyy-ww, yyyy-mm-dd (Warning) If it is known that a person died, it is usually expected that the DateOfDeath is reported.
24.DeathCause	DAIDS = Death due to AIDS; DNOAIDS = Non AIDS-related death; DUNK = Died of unknown cause (Remark) If death due to AIDS is reported, patient should also have a DateOfAIDSDiagnosis. (Warning) If DateOfDeath is known, DeathCause should be coded as DAIDS or DNOAIDS or DUNK. (Warning) If it is known that a person died, it is usually expected that the DateOfDeath is reported.

3.5 Data from the Portuguese Surveillance System

Among all the potential fields of Portuguese notification form, only 31 were made available. The variables collected until 2011 are described in Table 3.4. Since this data collection system is paper based, we were unable to classify the variables as “Mandatory” or “Optional”, so we chose to classify them as “Most likely to be complete” (**C**) or “Most likely to be missing” (**F**) based on the completeness level of each variable published by Mauch 2009 report [93] and on the requirements of the ECDC. Cases were considered ”most likely to be complete” whenever the percentage of ”blank” and ”unknown” responses was at least 25%.

Mauch refers that after data is manually entered, logical errors and inaccuracies are assessed and corrections made as needed. Validation rules are not routinely implemented by the system and inaccuracies are assessed by cross-tabulations. The identified format and values are represented in Table 3.5.

Table 3.4: Variables collected by the HIV-AIDS Portuguese notification form.

Variable	Report Type	Description
System Related Variables		
1. NO - Case Number	C	Number of the record in CVEDT
2. Soundex code	C	Encrypted small string patient identifier
3. Reception Date	C	Date of Reception at CVEDT
4. Clinical Stage	C	This variable specifies the clinical stage of the last notification
Diagnosis Information		
5. Notification Date	C	Date on which the case was notified for the last time.
6. Diagnosis Date	C	Date on which the case was diagnosed or progressed.
7. Notification HIV Date	C	Date on which the case was first notified.
8. Diagnosis HIV Date	C	Date on which the case was first diagnosed.
9. Stage Clinician	C	Clinical stage at the time of HIV diagnosis [93].
10. First symptoms	F	Date on which patient experienced the first symptoms
11. VIH 1	C	Was the patient infected with HIV 1 type of virus?
12. VIH 2	C	Was the patient infected with HIV 2 type of virus?
13. Risk Group	C	Most probable route of HIV transmission.

C: Most likely to be complete ; F: Most likely not to be filled (Table continues...)

Table 3.4: Variables collected by the HIV-AIDS portuguese notification form. - continue

Variable	Report Type	Description
Demographics		
14. Sex	C	Gender of the infected patient
15. Birth Date	C	Birth date of the patient
16. Age	C	Age at diagnosis of the person in years
17. Months	C	Age at diagnosis of the person in months
18. Nationality	C	Nationality.
19. Region of birth	F	Region of birth.
20. Residency	C	Region where the patient lives
21. Pregnant woman	F	In case of a female patient, whether or not she is pregnant
Clinical Information		
22. Year of Probable Infection	F	Probable year of infection
23. AIDS indicator disease 1	C	AIDS Indicator Disease
24. AIDS indicator disease 2	C	AIDS Indicator Disease
25. CD4	F	CD4 cell count at time of diagnosis.
26. CD8	F	CD8 cell count at time of diagnosis.
27. T4/T8	F	T4/T8 cell ratio at time of diagnosis (also called CD4 / CD8 cell ratio).
28. Treatment	F	Was the patient receiving antiretroviral therapy at the date last seen for care?
29. Hospital	C	Reporting Health Provider name
Death		
30. Death Date	F	Date of death due to any cause.

C: Most likely to be complete ; F: Most likely not to be filled

Table 3.5: Values and validation rules of the revised set of variables for case-based Portuguese HIV-AIDS surveillance

Variable	Codes, Allowed Formats and Validation Rules
System Related Variables	
1.NO	
2. Code	Until five capital letters and '-' characters
3. Reception Date	dd/mm/yyyy Reception date should always be filled (R25) Reception date must be after 1985 (R36) Reception date should be after Diagnosis date (5.) (R37) Reception date should be after Notification date (4.) (R39) Reception date should be after Diagnosis of HIV date (7.) (R38) Reception date should be after Notification of HIV date (6.) (R40) Reception date should be after Death date (30.)
4.Clinical Stage	PA = Asymptomatic Non-AIDS; SIDA = AIDS ; CRS = Symptomatic Non-AIDS (R31) If clinical stage of the last notification is reported as AIDS, then at least one AIDS indicator disease should be reported. (R57) If clinical stage of the last notification is reported as AIDS and patient has more than 5 years old, then CD4 should be less than 350 cells /mm^3. If AIDS patient has age less then 12 months then CD4 should be less than 30 cells / mm^3, if the age is between 1 and 2 years old should be less than 25 cells / mm^3 and with age between 3 and 4 should be less than 20 cells / mm^3.
Diagnosis Information	
5.Notification Date	ddd/mm/yyyyy (R26) The last notification date should be after 1985. (R41) The last notification date should be after diagnosis date. (R37) The last notification date should be before reception date. (R42) The last notification date should be after the notification of HIV date. (R43) The last notification date should be after the diagnosis of HIV date.
6. Diagnosis Date	dd/mm/yyyy (R27) The last diagnosis date should be after 1983. (R41) The last diagnosis date should be before last notification date. (R36) The last diagnosis date should be before reception date. (R44) The last diagnosis date should be after the diagnosis of HIV date.
7. Notification HIV Date	dd/mm/yyyy, NA (R28) The first notification date should be after 1985. The first notification date should be after first diagnosis date. (R38) The first notification date should be before reception date. (R42) The first notification date should be before last notification date. The first notification date should be before last diagnosis date.
8.Diagnosis VIH date	dd-mm-yyyy (R29) The first diagnosis date should be after 1983. The first diagnosis date should be before first notification date. (R39) The first diagnosis date should be before reception date.
9. Stage clinician	PA = Asymptomatic Non-AIDS; SIDA = AIDS ; CR = Symptomatic non-AIDS The first clinical stage should be at least as severe as the last clinical stage.
10. First symptoms	dd-mm-yyyy; NA (R45) The date of first symptoms should be before diagnosis date. (R46)The date of first symptoms should be before notification date.

(Table continues...)

Table 3.5: Overview of the revised set of variables for case-based HIV-AIDS surveillance - continue

Variable	Codes, Allowed Formats and Validation Rules
	(R47)The date of first symptoms should be before reception date. (R49) The date of first symptoms should be before first diagnosis date. (R48)The date of first symptoms should be before first notification date.
11.VIH 1	P = Positive; N = Negative; D = Unknown
12.VIH 2	P = Positive; N = Negative; D = Unknown
13.Risk Group	DADOR = Donor; DESCONHECIDO = Unknown; HEMOFILICO = Haemophiliac patient; HETEROSSEXUAL = Heterosexual contact; HOMO OU BISSEXUAL = MSM / homo or bisexual male; HOMO/TOXICODEPENDENT = Homo or bisexual male / Injected drugs; MAE/FILHO = mother-to-child-transmission; NOSOCOMIAL = Nosocomial patient; TOXICODEPENDENTE = Injected drugs; TRANSFUSIONADOS = Transfusion recipient; OUTRO/INDETERMINADO = Other or mined unknown. (R32) If risk group category is men who have sex with men, then sex should not be female. (R32) If risk group category is reported as transfusion of blood or its component due to haemophilia, then sex is not very often female. (R34) If risk group category mother to child transmission, then age is usually less than 13 years.
Demographics	
14.Sex	D = Unknown; F = Female; M = Male
15.Birth Date	dd/mm/yyyy and 11-11-1111
16.Age	Number from 0 to 99
17.Month	Number from 0 to 11
18.Nationality	Country name or nationality
19.Region of Birth	Districts, islands, autonomous regions (the Azores and Madeira), municipalities (concelhos), countries and continents
20.Residency	Districts, autonomous regions, countries, continents, NA
21. Pregnant Woman	N = No ; S = Yes (R33) If pregnant women is 'Yes', then sex must be female.
Clinical Information	
22.Year of Probable Infection	Four digit number, the number 9999 is also present meaning Unknown (R50) The year of probable infection should be before the year of last diagnosis. (R51) The year of probable infection should be before the year of last notification. (R52) The year of probable infection should be before the year of first symptoms. (R53) The year of probable infection should be before the year of first diagnosis. (R54) The year of probable infection should be before the year of first notification.
23. AIDS Indicator Disease 1	31 codes of AIDS indicator disease described in Appendix C. If AIDS indicator disease is filled then last clinical stage should be AIDS.
24. AIDS Indicator Disease 2	31 codes of AIDS indicator disease described in Appendix C. If AIDS indicator disease is filled then last clinical stage should be AIDS.
25.CD4	Numerical value (R15 and R16) Usually CD4 cell count varies in a range from 0 to 1,500 and due to rare extremely high values upper limit is set to 6,000 per cubic millimetre.
26.CD8	Numerical value (R17 and R18) Usually CD8 cell count varies in a range from 0 to 1,500 and due to rare extremely high values upper limit is set to 6,000 per cubic millimetre.
27.T4/T8	Numerical value (R35) Should be equal to the ratio between CD4 and CD8. (R19 and 20) Should be between 0.54 and 1.01 for middle-aged HIV-infected population on long term ART and with good immunovirological status [157].
28.Treatment	1 = Yes ; 2 = No; 9 = Unknown
29.Hospital	Name of Regional health administrations (RHAs); Hospital Name; Institutes of Ministry of Health; Ministry of Health Services; Primary Care Services Names; Addictive behaviour treatment centres; Private Practices; Districts ; Prisons; Others
Death	
30. Death Date	dd/mm/yyyy (R56) The death date should be before reception date. (R55) The death date should be before notification date

The Portuguese health care system has a very complex structure that suffered several reforms throughout time. In order to give a common understanding to the variable "Hospital" from the Portuguese surveillance system, this variable was cross-checked with the official list of Health Providers made available by the Portuguese Ministry of Health in Website of Central Administration of the Health System (ACSS). The description of these variables is presented in Table 3.6.

Table 3.6: Central Administration of the Health System List of Health Providers

Variable	Description	Values
ARS	Health Administration Region	ARS Alentejo; ARS Algarve; ARS Centro; ARS LVT; ARS Norte
COD LOCAL PRESCR	Code of Health Provider	Numerical value
NOME LOCAL PRESCR	Name of Health Provider	A character
TIPO LOCAL	Type of Health Provider	Primary Care; Personalized Care; Diagnosis Centre; Private Care; Hospitals; Shared Services; Counseling Centre; ...
RESP FINANCEIRA	Name of Financial Manager	ARS Alentejo; ARS Algarve; ARS Centro; ARS LVT; ARS Norte; Personalized Health Care Unit
DATA INICIO RESP FINANCEIRA	Beginning date of Financial Manager	

(Table continues...)

Table 3.6: Central Administration of the Health System List of Health Providers - continue

Variable	Description	Values
DATA FIM RESP FINANCEIRA	Late date OF Financial Manager	
CONCELHO	Region	Portuguese Region
ACTIVO	If it is active	
DATA INICIO ARS	Beginning date of administrative responsibility	
DATA FIM ARS	Last date of administrative responsibility	

As previously mentioned, the Portuguese HIV-AIDS reporting system is confidential, and case- and paper-based. It was performed voluntarily by physicians since the identification of the infection in 1983 and until 2005. Since then, it is mandatory to report all cases in any of the stages of the infection (asymptomatic, AIDS related complex and AIDS) and all the progressions including death. The stages of the infection are determined according to case definition of WHO and the ECDC since 1993. The registry also includes information on the most probable source of transmission (Heterosexual, Men who have Sex with Men and Injecting Drug Users), as well as demographic and clinical variables. The demographic variables include age, sex, birth date, nationality and region of birth. The diagnosis information includes notification date, diagnosis date, notification of HIV date, diagnosis of HIV date, year of the first symptoms and type of HIV virus. Clinical Information includes probable year of infection, codes of AIDS Indicator Disease, CD4 and CD8 cells count, T4/T8 cell ratio at the time of diagnosis, the status of the treatment and the name of the reporting Health Provider. The death information includes date of death due to any cause. Patients confidentiality is assured through the use of a Soundex code filled in by the clinicians and data is processed manually.

It is important to notice that the entire Portuguese population is virtually under surveillance but the testing practices are mandatory for military and strongly recommended for pregnant women and IDUs entering treatment programs.

3.5.1 Evaluation of the HIV-AIDS Portuguese Surveillance Data

In 2009, the Portuguese HIV-AIDS reporting system situation was assessed and the results were published in [93]. By then, it was acknowledged, like in most similar surveillance systems, that the Portuguese surveillance suffers from under-reporting, under-diagnosing and reporting delay, clearly affecting timeliness and therefore the data quality. The study aimed to examine the quality of the data collected by the portuguese mandatory notification system for HIV - AIDS infection as determined by the timeliness of reporting, completeness of information, acceptability by relevant stakeholders, as well as simplicity, stability and flexibility of the surveillance system.

Mauch found that the acceptability, as measured through qualitative interviews with key stakeholders, is relatively low. These interviews were conducted on a selection of health providers that play a key role in HIV - AIDS care and treatment in seven different hospitals in Lisbon and Porto. These seven hospitals have been responsible for reporting 41% of all cases notified. All clinicians interviewed were either the Director of the Infectious Disease Service or a practising physician on the service and so are important in the notification process. Five clinicians regularly see 12 to 20 patients per day while the other follow only a handful of patients on a regular basis. The hospitals varied on the number of HIV+ patients being followed in their Infectious Disease Service but ranged from 1300 to 2500 patients [93]. Also, she found that no standard operating procedures, such as description of objectives and responsibilities of an institution operating within the system, were implemented until the time of the study [93].

All clinicians agree that the system is useful, confirmed knowing that the notification is mandatory and under what conditions regarding clinical stage, confirmed that the submission was preformed on a regular basis for initial diagnosis. The main reasons for not submitting a notification form are related with paperwork not representing a priority, forget, consider that the infra- structure of the process is difficult to deal with, overwork and the necessity of large backlog of patient records to review. Clinicians did not report being accountable to any time constraints. This leaves a considerable amount of leeway for health providers as to how quickly they should be submitting a notification form. They reported that the filling in of the notification form was occurring after the medical appointment, sometimes several days or weeks later, also referred that there were some personal judgement about the importance of filling some variables, that do not have “proprietary” ownership of the data and that some patients had problems recalling some events. This fact can contribute to missing values and inaccuracies in some variables such as the associated risk group[93]. According to these key stakeholders, reporting death and changes in disease status are particular problematic.

This informations are summarized in Table 3.7.

Table 3.7: Summary highlights of Health Provider Interviews. Adapted from Mauch in [93]

	Hospital A	Hospital B	Hospital C	Hospital D	Hospital E	Hospital F	Hospital G
View the Notification System as useful	Agree	Agree	Agree	Agree	Agree	Agree	Agree
Received training or support regarding data submitted via notification	Disagree	Disagree	Agree	Disagree	Disagree	Disagree	Disagree
Receive regular reports from the CVEDT	Disagree	Disagree	Agree	Agree	Disagree	Agree	Agree
Regularly receives confirmation from CVEDT that notifications were received	Agree	Agree but coded so difficult to compare to internal records	Disagree	Agree	Agree but coded so difficult to compare to internal records	Disagree	Disagree
Case definition being used to classify disease status of patient	1993 CDC	1993 CDC, minus CD4 cell count criteria	1993 CDC	Unsure	1993 CDC	ECDC	1993 CDC, minus CD4 cell count criteria
Clinicians individually responsible for submitting notification forms	Agree	Agree	Agree	Disagree - Centralized to 2 doctors	Agree	Agree	Disagree - Centralized to 1 doctor
Aware that notification is mandatory at first diagnosis and any change in disease status, including death	Agree	Agree	Agree	Agree	Agree	Agree	Agree
Regularly submits notification for initial diagnosis	Agree	Agree	Agree	Agree	Agree	Agree	Agree
Regularly submits notification for disease progression in patient (PA to CRS, CRS to AIDS)	Disagree	Disagree	Disagree	Disagree	Disagree	Reports change to AIDS status, not to CRS	Disagree
Regularly submits notification for death of patient	Disagree	Disagree	Disagree	Disagree	Disagree	Disagree	Disagree
Reason for not submitting a notification form	Paperwork not a priority	Paperwork not a priority, Forget	Infra- structure of the process makes it difficult	Large backlog of patient records to review	No reason given	Overwork, Forget	No reason given
Time required to complete notification form	2 minutes for regular patient, up to 10 min for new patients	10 minutes	10 minutes for regular patient, up to 30 min for new patients	10 minutes for own patient, up to 1 hour for patient of other doctor	2 minutes for regular patient, up to 10 min for new patients	3 to 4 minutes	5 to 10 minutes
Variables most likely to not be completed	Motive to test, Disease classification	Risk group, concelho, Country of probable infection, Date for first HIV+ test	No specific variables	Foreign travel, Date for first HIV+ test	Motive to test, Foreign travel	Motive to test, Military service	Risk group
Reasons for not completing a variable	Not relevant to CVEDT needs	Information not requested, Patient recall	Lab results not available, Patient recall	Information not requested, Too time- consuming	Not relevant to CVEDT needs	Information not requested, Patient recall	Patient recall
Support implementation of electronic notification system	Disagree	Agree	Agree	Agree	Agree	Agree	Agree
Feasible to add variables already in use for clinical purposes (i.e. CD4, treatment)	Agree	Agree	Agree	Agree	Agree	Agree	Agree

Over the years, the surveillance procedure has suffered some changes that may have altered the quality of the reports. One important change was the inclusion of relevant new variables in 1988 due to the evolution of knowledge about the natural history of the infection

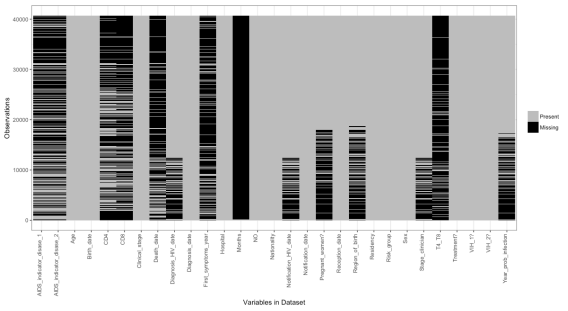


Figure 3.3: Missing values per variable

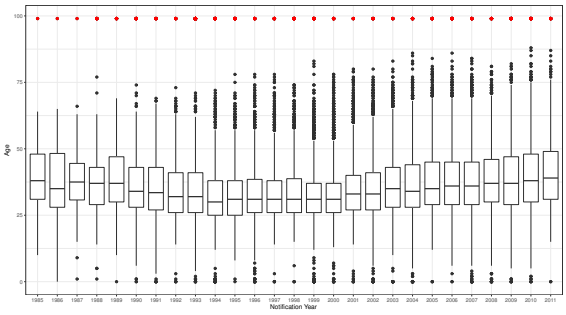


Figure 3.4: Distribution of age by notification year

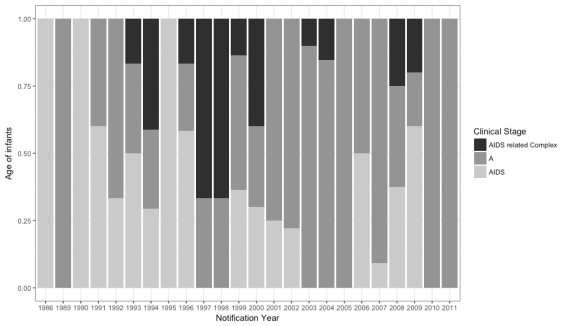


Figure 3.5: Distribution of age of babies by clinical Stage

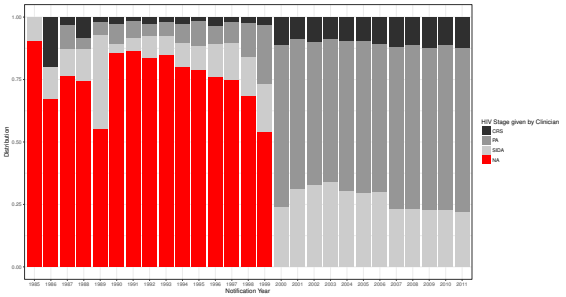


Figure 3.6: Distribution of the HIV stage reported by the clinician.

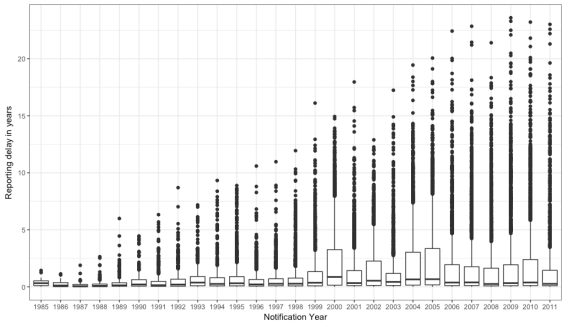


Figure 3.7: Distribution of reporting delay per notification year

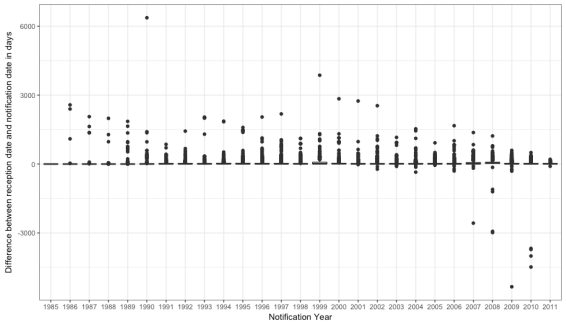


Figure 3.8: Distribution of time between reception and notification of the form at CVDET per notification year

and the elimination of others in 2005 [93].

One of the first tasks consisted of determining the number of missing cases on each variable. Figure 3.3 elucidates about the global picture happening in the dataset with respect to the number of missing cases in each variable and the identification of the cases mostly affected by missing information. Twelve variables presented more than 20% of missing values. However, in the variables *Pregnant-women*, *AIDS indicator disease 1*, *AIDS indicator disease 2* and *Months*, most of the missing values were structural and, although the variable *Region of Birth* appears to be non - missing, most of the observations are *Not Mentioned* (about 11958 cases).

The text fields *Nationality*, *Region-of-birth* and *Hospital* have issues such as misspelling, illegal values, ambiguities and syntax violations (Table 3.8).

Table 3.8: Some issues for variable “Nationality” and “Hospital”

Error	Type	Frequency
PORTGUESA	Misspelled by deletion of character	1
PORTUGUESAA	Misspelled by insertion of a character	1
PORTUGAL	Illegal Value	13
BRASILEIRO	Misspelled by substitution of a character	2
Univ Coimbra	Syntaxe violation and ambiguous	

The most common ambiguities are values representing names of Administrative Health Regions, names of Districts and short names of Health Care Institutions as the one mentioned in Table 3.1.

When the distribution of age according to the year of notification was looked at (Figure 3.4), it became clear that the very suspicious value “99” did not correspond to years of age but to a code form for a missing value instead.

The interpretation of positive HIV antibody testing is complicated in children less then 9−12 months due to the fact that maternal HIV antibody can persist for 18 months and so, the standard HIV testing cannot be reliable. The distribution of age of babies under 9 months - old by clinical stage was assessed and it is present in Figure 3.5.

We defined missing values in any variable as the entries with blank fields and coded as missing values such as, “9”, “99”, “9999”, “Not Mentioned”, “Unknown”,“D” (first letter of Unknown in Portuguese), “Other/Undetermined” or “——”, “—(any two letters)”, “(any three letters)–” or “AAAAA”. The entries with illegal values, wrong data or misspellings were considered as invalid values.

The results of HIV-AIDS data audit are presented in Table 3.9 and graphically in Figure 3.9. From the analysis of these elements it is clear that almost 50% of the variables have potential for significant data quality problems since they are severely affected by missing values. The lack of information is sharp on variables like “Months”, “CD4”, “CD8”,“T4/T8”, “Region of Birth”, “Treatment”, “AIDS indicator disease 1”, “AIDS indicator disease 2”, “First symptoms” , “Death date” and “Year of probable infection”. It is also of note that the variable “Hospital” has a moderate amount of invalid values.

In Table 3.9, the number it is also available the the number of distinct observation each may influence the representativeness of some groups, this is especially problematic in the variable “Hospital”.

Table 3.9: Percentage of Valid, Invalid, Missing and Outliers

Name	Valid	Invalid	Missing	Outliers	Distinct
ID Variable — 2 Variables 40690 Observations					
NO- Case Number	100%	0%	0%		
Numerical Variables — 5 Variables 40690 Observations					
Age	95%	0%	1%	4%	89
Months	0%	0%	100%		12
CD4	40%	0%	59%	1%	1355
CD8	29%	0%	70%	1%	2390
T4/T8	14%	0%	85%	1%	232
Factor Variables — 14 Variables 40690 Observations					
Clinical stage	100%	0%	0%		3
Sex	100%	0%	0%		2
Pregnant women	64%	0%	36%		2
Stage clinician	79%	0%	21%		3
Region of birth	36%	3%	61%		445
Residency	98%	0	2%		24
Nationality	94%	1%	5%		101
Risk group	98%	0%	2%		9

(Table continues...)

Table 3.9: Percentage of Valid, Invalid, Missing and Outliers

Name	Valid	Invalid	Missing	Outliers	Distinct
Treatment	8%	0%	92%		2
AIDS indicator disease 1	41%	0%	59%		30
AIDS indicator disease 2	11%	0%	89%		27
VIH 1	99%	0%	1%		2
VIH 2	99%	0%	1%		2
Hospital	86%	14%	0%		521
Dates Variables — 9 Variables		40690 Observations			
Reception date	100%	0%	0%		4510
Birth date	99%	0%	1%		15301
Notification date	100%	0%	0%		6065
Diagnosis date	100%	0%	0%		7073
Notification HIV date	79%	0%	21%		4961
Diagnosis HIV date	79%	0%	21%		6052
First symptoms	26%	0%	74%		2426
Death date	23%	0%	77%		5256
Year probable infection	16%	0%	84%		41

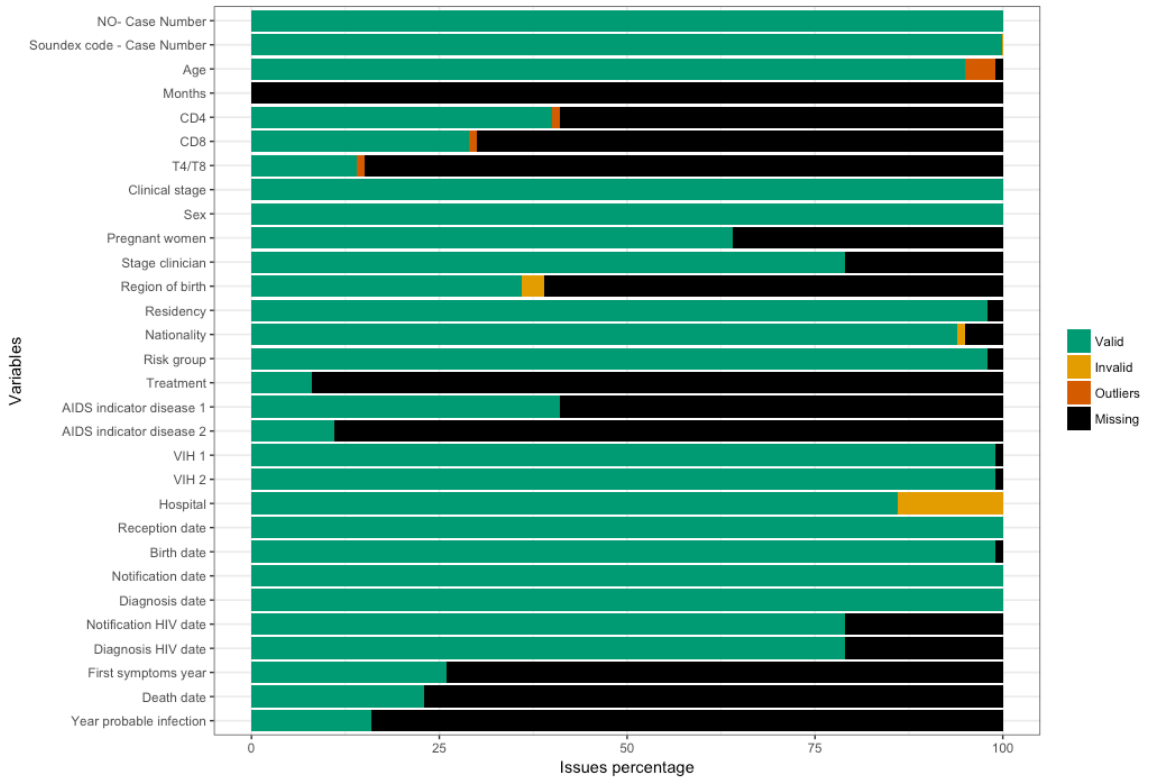


Figure 3.9: Percentage of Valid, Invalid, Missing and Outliers

Throughout the years, the surveillance system suffered changes that are imprinted in the data. One example includes the variable “HIV clinical stage reported by the clinician”, as seen in Figure 3.6.

The presence of reporting delay and longitudinal outliers is represented in Figure 3.7. It can also be seen that there are reporting delays longer than 25 years and that this data has a visible right truncation mechanism.

In any process of data editing it is necessary to take into consideration all the relationships between the variables in the data set. The constraints describing the relationships are called *functional dependencies*. Given the biology of the HIV-AIDS disease it is natural to define the dependency between the “CD4” and “CD8” cell counts and as a consequence

of the ratio between them, which is commonly represented by the “T4_T8” ratio. Moreover, the variable “Sex” is related to “Risk group” due to the transmission patterns of the disease and linked to the variable“Pregnant woman”. Another natural relationship occurs between “Age” and “Birth date” variables. Moreover, given the natural history of the disease and the chronology of each event, all variables concerning dates are related. Since the “Clinical Stage” is defined by clinical and laboratory criteria it is natural that these variables are related to “CD4”, “CD8”, “T4_T8”, “AIDS_indicator_disease1”, “AIDS_indicator_disease2” and “Age”. The functional dependencies among all variables are represented in Figure 3.10.

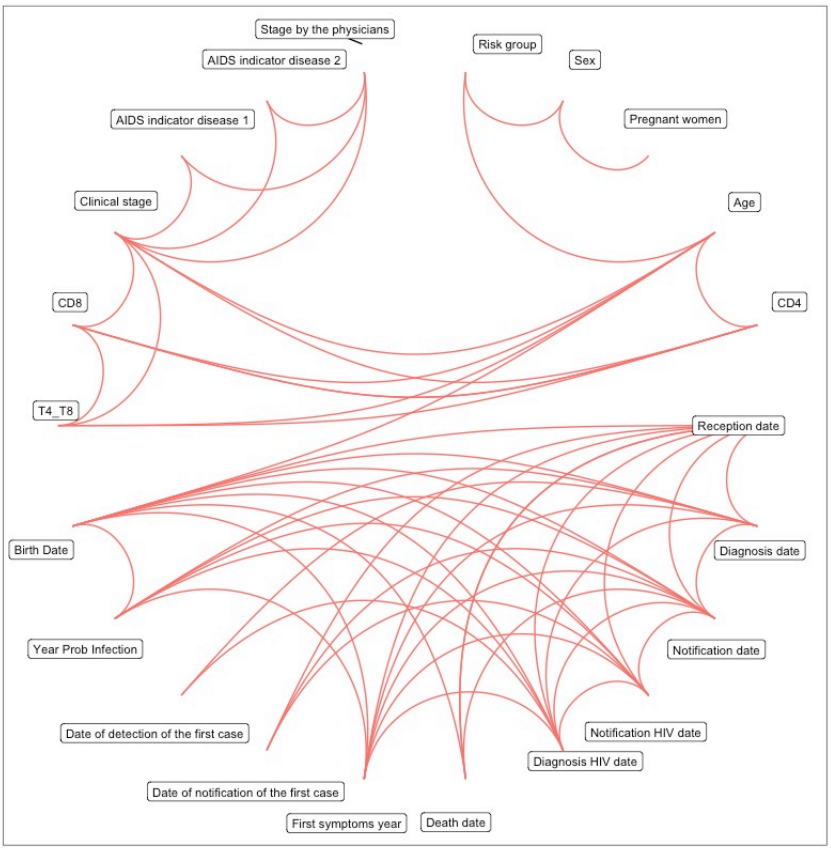


Figure 3.10: Functional dependencies among Portuguese HIV-AIDS surveillance system variables

Considering these relationships and applying the business rules defined in Table 3.5 more inconsistencies appeared.

Violation of logical dependencies are also presented in time related variables. For example, when considering the relation between the reception of form in CVEDT date (filled by CVEDT) and notification date (filled by clinician), recent dates can be found in the latter (Figure 3.8).

The validation rules in Table 3.10 were deducted based on the rules implemented on TESSy for the HIV - AIDS Surveillance System and extended for accommodating some logical or mathematical relationships and are present in Table 3.5. So, the completeness property is satisfied. The initial set was pruned in order to eliminate redundancies and inconsistencies.

The final set of validation rules is described in terms of material conditions and evaluated in Table 3.10 and Figure 3.11. Theses rules were applied to Portuguese HIV data set. All rules involving “Death date”, “Year Probable Infection”, “Treatment”, both “AIDS Indicator disease”, “Months”, “CD4”, “CD8” and the ratio ”T4_T8” are severally affected by

missing values. The missing values of variable “Months” are due to the nature of the disease which affects mainly young adults and due to HAART vertical transmission is almost null in Portugal. Some lack of information on “Treatment”,“CD4”, “CD8” and the ratio ”T4_T8” is due to the 2005 revision: these variables were excluded from the notification form ([93]). “Nationality” and “Hospital” were the variables counting with more invalid or ambiguous values, due to typo’s or values representing other statistical objects. It is important to notice that the rule number R57 takes into consideration the laboratory and clinical criteria to AIDS stage definition.

Table 3.10: Validation rules applied to HIV-AIDS Portuguese data set

Rule	Expression	% Confirmed	% Failed	% Missing
R1	Risk group $\in$ {”Dador” , ”Desconhecido” , ”Hemofilico” , ”Heterossexual” , ”Homo ou Bissexual” , ”Homo/Toxicodependent” , ”Mae/Filho” , ”Nosocomial” , ”Outro/Indeterminado” , ”Toxicodependente” , ”Transfusionado”}	98%	0%	2%
R2	Sex $\in$ {”F”,”M”}	100%	0%	0%
R3	Pregnant women $\in \{”Yes”, ”No”\}$	64%	0%	36%
R4	Clinical stage $\in \{”PA”, ”SIDA”, ”CRS”\}$	100%	0%	0%
R5	Stage Clinician $\in \{”PA”, ”SIDA”, ”CRS”\}$	79%	0%	21%
R6	Region of birth	36%	36%	32%
R7	Residency	98%	0%	2%
R8	Nationality	94%	1%	5%
R9	Treatment	8%	0%	92%
R10	AIDS indicator disease 1	41%	0%	59%
R11	AIDS indicator disease 2	11%	0%	89%
R12	$VIH1 \in \{P, N\}$	99%	0%	1%
R13	$VIH2 \in \{P, N\}$	99%	0%	1%
R14	Hospital	86%	14%	0%
R15	CD4 >0	41%	0%	59%
R16	CD4 <6000	41%	0%	59%
R17	CD8 >0	30%	0%	70%
R18	CD8 <6000	30%	0%	70%
R19	(T4_T8 - 0.54) >= -1e-08	4%	11%	85%
R20	(T4_T8 - 1.01) <= 1e-08	14%	1%	85%
R21	(Age - 0) >= -1e-08	99%	0%	1%
R22	Age <98	99%	0%	1%
R23	(Months - 0) >= -1e-08	0%	0%	100%
R24	Months <12	0%	0%	100%
R25	Reception date >= ”1/1/1985”	100%	0%	0%
R26	Notification date >= ”1/1/1985”	100%	0%	0%
R27	Diagnosis date >= ”1/1/1983”	100%	0%	0%
R28	Notification HIV date >= ”1/1/1985”	79%	0%	21%
R29	Diagnosis HIV date >= ”1/1/1983”	79%	0%	21%
R30	Death date >= ”1/1/1983”	23%	0%	77%
R31	!(Clinical stage == ”SIDA”) | ((is.na(AIDS_indicator_disase_1) == FALSE | is.na(AIDS_indicator_disase_2) == FALSE))	100%	0%	0%
R32	!(Risk_group %in% c(”Homo/Bissexual”, ”Homo/Toxicodependent”, ”Hemofilico”)) | (Sex == ”M”)	100%	0%	0%
R33	!(Pregnant_women == ”S”) | (Sex == ”F”)	72%	0%	28%
R34	!(Risk_group == ”Mae/Filho” & Clinical_stage =“PA”) | (Age_first_diagnosis <= 13)	100%	0%	0%
R35	round(CD4/CD8, digits =2) == T4_T8	7%	6%	87%
R36	Reception_date >= Diagnosis_date	100%	0%	0%
R37	Reception_date >= Notification_date	100%	0%	0%
R38	Reception_date >= Notification_HIV_date	79%	0%	21%
R39	Reception_date >= Diagnosis_HIV_date	79%	0%	21%
R40	Reception_date >= Death_date	11%	11%	77%
R41	Notification_date >= Diagnosis_date	100%	0%	0%
R42	Notification_date >= Notification_HIV_date	79%	0%	21%
R43	Notification_date >= Diagnosis_HIV_date	79%	0%	21%
R44	Diagnosis_date >= Diagnosis_HIV_date	79%	0%	21%
R45	!(Clinical_stage != ”PA”) | (First_symptoms_year <= Diagnosis_date)	72%	0%	28%
R46	!(Clinical_stage != ”PA”) | (First_symptoms_year <= Notification_date)	72%	0%	28%
R47	!(Clinical_stage != ”PA”) | (First_symptoms_year <= Reception_date)	72%	0%	28%
R48	!(Clinical_stage != ”PA”) | (First_symptoms_year <= Notification_HIV_date)	63%	0%	36%
R49	!(Clinical_stage != ”PA”) | (First_symptoms_year <= Diagnosis_HIV_date)	63%	1%	36%
R50	Year_prob_infection <= Diagnosis_Year)	16%	0%	84%

(Table continues...)

Table 3.10: Validation rules applied to HIV-AIDS Portuguese data set

Rule	Expression	% Con-firmed	% Failed	% Miss-ing
R51	Year_prob_infection <= Notification_Year	16%	0%	84%
R52	Year_prob_infection <= First_symptoms_Year	3%	0%	97%
R53	Year_prob_infection <= Diagnosis_HIV_Year	16%	0%	84%
R54	Year_prob_infection <= Notification_HIV_Year	16%	0%	84%
R55	Death_date <= Notification_date	11%	12%	77%
R56	Death_date <= Reception_date	11%	11%	77%
R57	(!((Age >= 5 & CD4 < 350) | is.na(AIDS_indicator_disase_1) == FALSE) | (Clinical_stage == ”SIDA”)) & (((Age >= 5 & CD4 < 350) | is.na(AIDS_indicator_disase_1) == FALSE) | (((!((Age < 1 & CD4 < 30) | is.na(AIDS_indicator_disase_1) == FALSE) | (Clinical_stage == ”SIDA”)) & (((Age < 1 & CD4 < 30) | is.na(AIDS_indicator_disase_1) == FALSE) | ((!((Age > 0 & Age < 3 & CD4 < 25) | is.na(AIDS_indicator_disase_1) == FALSE) | (Clinical_stage == ”SIDA”)) & (((Age > 0 & Age < 3 & CD4 < 25) | is.na(AIDS_indicator_disase_1) == FALSE) | (!((Age > 2 & Age < 5 & CD4 < 20) | is.na(AIDS_indicator_disase_1) == FALSE) | (Clinical_stage == ”SIDA”)))))))	54%	8%	38%
R58	lubridate::year(Birth_date) <= Year_prob_infection	0%	16%	84%
R59	Birth_date <= First_symptoms_year	25%	0%	75%
R60	Birth_date <= Death_date	23%	0%	77%
R61	Birth_date <= Diagnosis_HIV_date	79%	0%	21%
R62	Birth_date <= Diagnosis_date	99%	0%	1%
R63	Birth_date <= Notification_HIV_date	79%	0%	21%
R64	Birth_date <= Notification_date	99%	0%	1%
R65	Birth_date <= Reception_date	99%	0%	1%

It is important to notice that R35 is severally impaired due to missing values in at least one of the 3 variables that composes this relationship and due to violation of the functional relationship (Figure 3.11). Indeed, variable “T4_T8” have a lot of inconsistencies, some of which are presented in Table 3.11.

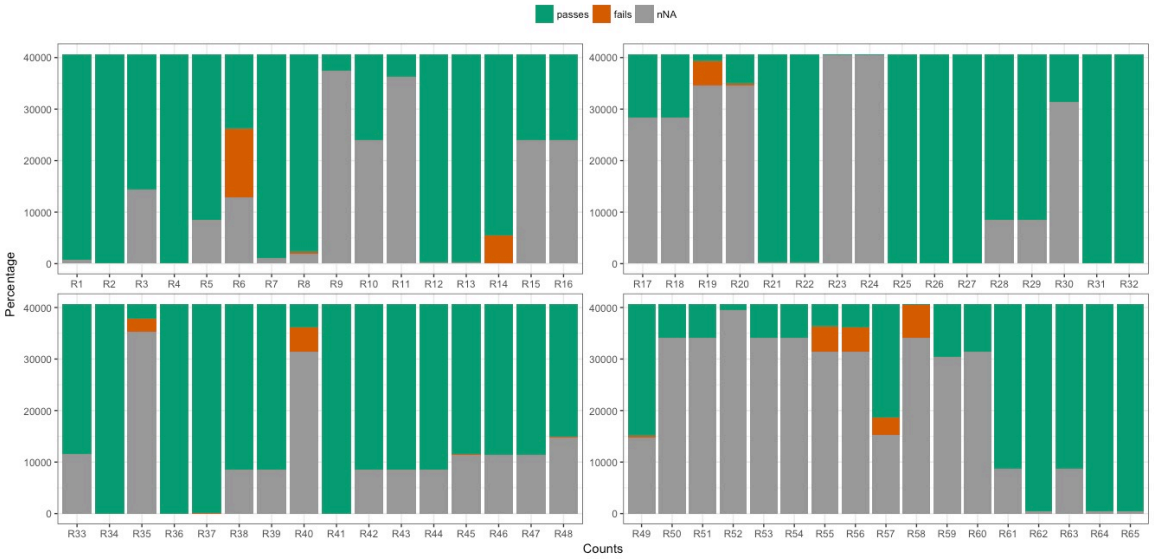


Figure 3.11: Validation rules applied to HIV-AIDS Portuguese data set

Other important remark is that the AIDS case definition changed along the studied surveillance years and in this study the data were confronted with the 2012 European AIDS case definition and so the failures reported by rule 57 may be roots on the evolution of the system.

Table 3.11: Examples of inconsistencies on T4_T8 records

CD4	CD8	CD4/CD8	T4_T8
102	318	0.32	3.20
286	638	0.45	4.50
19	236	0.08	8.00
36	24	1.5	1.50
952	418	2.30	2.30
389	525	0.74	1.90
57	314	0.18	6.00
42	449	0.09	0.01

When this methodology is applied from the record perspective it can be seen that all the records have some percentage of rules that can not be evaluated due to missing values in variables involved on the rule definition. The majority of the records have at least 25% and a maximum of 62% of rules that can not be evaluated. It can be found at least a record with a maximum of 25% of failed rules and a record with just 25% of the rules confirmed 3.12. After correcting the trivial typo's in the data set and given the set of validation rules which takes into consideration the web of dependent variables given in Figure 3.10 so that the correction of a record in order to repair a rule does not make it invalid for other rule, a record based error detection framework was implement. This framework implements the Felligi Holt algorithm minimizing a weighted number of values that need to be adjusted to remove the invalidation. It was used the *errolocate* R package which translates the validation and error localization problem into a mixed integer problem and uses - a Mixed Integer Programme solver to find a solution [145, 158]. Taking into consideration the total percentage of failed or missing values and given that some variable have some degree of uncertainty due to recall errors and stigma.

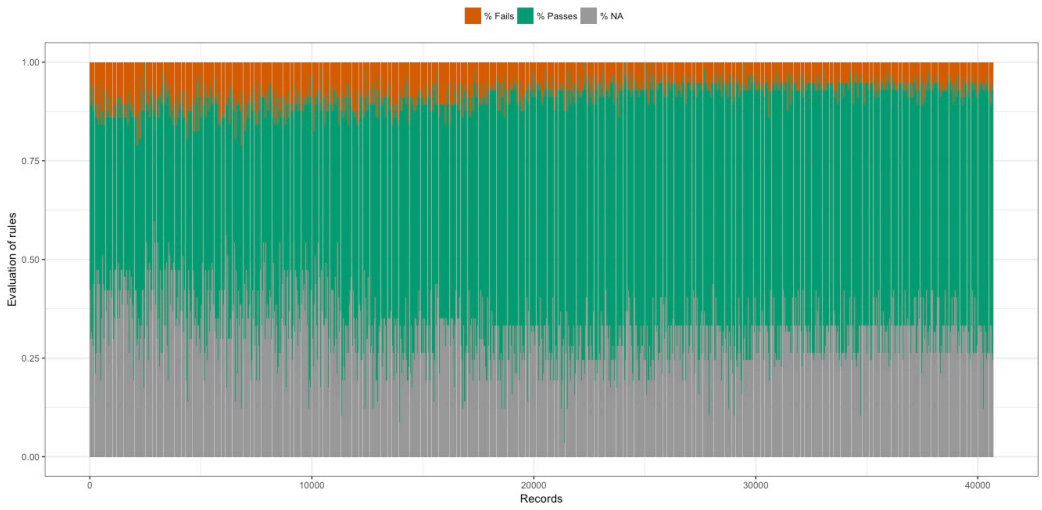


Figure 3.12: Evaluation of validation rules applied to HIV-AIDS Portuguese data set per record

It was also considered that the "Diagnosis date" has some uncertainty in it. The variables were ordered and weights were given accordingly, with a low weight corresponding to a higher uncertainty [159, 145]:

W1 T4_T8;

W2 Death_date;

W3 Year_prob_infection, First_symptoms_year, Treatment, CD8 ;

W4 Region_of_birth, CD4;

W5 Diagnosis_HIV_date ;

W6 Notification_HIV_date, Pregnant_women, Stage_clinician, Nationality, Residency;

W7 Diagnosis_date; Risk_group ;

W8 Notification_date, AIDS_indicator_disease_1 , AIDS_indicator_disease_2, VIH_1, VIH_2, Hospital

W9 Sex, Birth_date, Age, Months ;

W10 Clinical_stage, Reception_date

This statistical cleaning step revealed that almost 99% of the records did not violate the logical dependencies between the variables variables, 5123 records needed a change in one variable and 17 records needed a change in two variables (Table 3.12).

Table 3.12: Summary of Error Location Algorithm per Record

Number of Errors	Number of Records
0	35550
1	5123
2	17

Table 3.13: Summary of Error Location Algorithm per Variable

Variable	Number of issues	Missing
T4_T8	5088	34544
CD4	36	23994
CD8	28	28357
Risk_group	1	0

It is clear that the percentage of missing values in each variable is an important issue and it was found that problems in laboratory variables must also be taken into consideration. The framework determined that the minimal set of variables needing an adjustment consists of the variables “T4_T8”, “CD4”, “CD8”, “Risk_group” and “Age at the first diagnosis” with errors and missing values represented in Table 3.13.

The detected issues on the variable “Risk_group” were not corrected since the rule represents only a soft edit.

Table 3.11 suggested that some of the values were records with some typographical errors. The records with valid values of “CD4” and “CD8” were used for proxy impute the “T4_T8” variable.

The duplicates or record linkages arised, most probably, by one of the following mechanisms: a patient seeked care on a primary care centre and, after confirmation of the infection, the patient is transferred to a specialized unit; a hospitalized patient developed a disease that needs special treatment offered in a specific unit; patient moved to a different region and, once there, seeked for a new health centre; health care; patients that were in prisons and had their HIV positive status confirmed need to be followed by an hospital; changes in disease stage and possible errors on key identifier and changes on the system dataset structuring. For duplicate detection, it would be necessary to have a meaningful observation identifier but the original Soundex Code could not be provided.

Given the high proportion of missing values in several variables, the presence of typographical errors and invalid values, we adopted an iterative deterministic technique for detecting record linkage. First we considered the variables “Encrypted Code”, “Sex” and “Birth_date”. For the generation of the comparison patterns it was established that the records should exactly match on the variables “Sex” and “Birth_date” and have, at least, a similarity of 0.9 with ‘Encrypted Code”. We used the Jaro similarity measure, taking values

in $[0, 1]$, where 0 denotes maximal dissimilarity and 1 stands for equality. This allows for 'fuzzy' comparison patterns and is useful for dealing with typographical errors and with small length strings. We allowed misspell errors in identifiers adjusting the distance in order to accommodate one character deletion, one insertion, one substitution and one transposition.

Given two strings $s = a_1 \dots a_K$ and $t = b_1 \dots b_L$, a character a_i in s is said to be common with t if there exists a j such that $a_i = b_j$ and $|j - i| < \frac{min(|s|,|t|)}{2}$, where $|s|$ (resp. $|t|$) is the number of characters forming the corresponding string [160, 155, 161].

Let $s' = a'_1 \dots a'_K$ be the characters in s that are common to t (exactly in this order) and let $t' = b'_1 \dots b'_L$ be an analogous string. A transposition for s', t' can be defined as a position i such that $a'_i \neq b'_i$. Let T be half the number of transpositions necessary to turn s' into t' [161] . The Jaro similarity measure for s and t is

$$d_{jaro}(s,t) = \frac{1}{3}\left(\frac{|s'|}{|s|} + \frac{|t'|}{|t|} + \frac{|s'| - T}{|s'|}\right) \tag{3.9}$$

A variation of this measure is the Jaro-Winkler measure that uses the length P of the longest common prefix of s and t. This measure incorporates a penalty for characters mismatches in the first four characters and so, it favourables pairs of strings containing identical prefixes. Letting $P' = \max(P, 4)$, the Jaro-Winkler measure is defined as [161]

$$d_{jaro-winkler}(s,t) = d_{jaro}(s,t) + \frac{P'}{10}(1 - d_{jaro}(s,t)) \tag{3.10}$$

A record pair is classified as a match if the two records agree, character by character, on all identifiers and the record pair is uniquely identified (no other record pair matched the same set of values). A record pair is classified as a non match if the two records disagree on any of the identifiers, if the record pair is not uniquely identified or if there is reasonable doubt on the match status (this dataset undergoes a revision when a new case is entered by a trained CVEDT employee so if there is a doubt about the matching status of two records, then we decide by a non match status) [152].

In the studied dataset, the approach generated 352299493 comparison pairs from which 280 pairs were classified as possible matches. A second step, with deterministic approach with clerical review was performed considering all the available information. Given that the variables have a considerable amount of missing values, it can be defined the following rules:

R1 For each pair, does one element come from a primary care health provider or association and the other come from a specialized health care provider? If so, the pair can represent the same patient

R2 For each pair with different death dates, is the difference larger than one month? If so, the pair can correspond to different patients.

- **R2.1** If just one has a death date, does the other the other have a later date of diagnosis? If so, the pair may correspond to different patients.

R3 Comparing the diagnosis dates, does the most recent corresponds to a latter stage of the disease? Then it can be different patients.

R4 Is the case from a minority group (ex: Cabo-Verde, Angola, Açores and Homo-Bissexual, Female and IDUs, ...)? Then probably is the same case.

R5 Given that the two notifications came from different hospitals, if the region of residency is on the second Hospital region, then it can be the same case.

R6 Given two AIDS cases, does the AIDS indicator diseases the same? If they have the same AIDS indicator disease then it is probably the same patient.

R7 Given that the two AIDS case, does the AIDS indicator disease(s) need a special care? Verify if the health provider is compatible with the special need.

R8 Given that the cases are a women, compare if the cases are pregnant. If yes, then is very likely that is the same case.

The risk group was not considered as a possible match field due to the existence of overlapping risk groups and the widely recognized uncertainty in the classification of the patients. Examples of the resulting matched pairs are presented in Table 3.14.

Table 3.14: Examples of matched pairs

Clinical stage	Sex	Birth date	Notification date	Diagnosis date	Notification HIV date	Diagno HIV date	Stage clinician	First symptoms year	Year prob infection	Region of birth	Residency	Pregnant women	AIDS indicator disease 1	Risk group	Hospital
A	F	04/05/1974	02/02/1996	25/02/1995	NA	NA	NA	NA	NA	NA	Lisboa	NA	NA	Toxicodependente	C.S. Lapa
A	F	04/05/1974	03/02/2011	29/10/1997	03/02/2011	29/10/1997	PA	NA	1997	Lisboa	Lisboa	N	NA	Heterossexual	Santa Maria
AIDS	F	01/03/1955	01/03/2002	30/11/2001	01/03/2002	30/11/2001	AIDS	15/11/2001	NA	NA	Lisboa	N	Candidiasis, oesophageal	Heterossexual	Desterro
AIDS	F	01/03/1955	15/02/2002	20/11/2001	15/02/2002	20/11/2001	AIDS	NA	NA	NA	Lisboa	N	Candidiasis, oesophageal	Heterossexual	Capuchos
A	F	31/07/1951	26/03/2008	25/03/2008	26/03/2008	25/03/2008	A	NA	NA	Porto	Porto	N	NA	Heterossexual	Sao Joao
A	F	31/07/1951	28/03/2008	24/03/2008	28/03/2008	24/03/2008	A	NA	NA	NA	Porto	N	NA	Heterossexual	Sao Joao

From the 280 pairs identified previously, only 132 were considered as record linkage in the second approach. It is important to notice that the coding variable have missing values indicated by the codes "—" and "AAAAA". These records were classified as non matches.

For the analysis of the missing values patterns, the observations were ordered per notification year Figure 3.13. It can be seen that the variables "Stage_clinician", "Diagnosis_HIV_date" and "Diagnosis_HIV_date" were most frequently missing in the early years while the variables "CD4", "CD8" and "T4_T8" presented a higher concentration of missing values in the most recent years. It is important to notice that these variables were dropped in the 2005 revision. Most of the missing values present in the variables "Pregnant_women", "AIDS_indicator_disease_1" and "AIDS_indicator_disease_2" are structural because "Male" is the prevalent sex and all cases classified as "AIDS" have at least one indicator disease.

Considering only the variables with missing values and given that the trivial missing cases on "pregnant_women" and "T4_T8" were corrected, it can be seen that when one variable has a missing value than all the variables are missing (Figure 3.14).

As already said, over the years, the portuguese HIV/AIDS surveillance system has suffered several modifications. This changes are highlighted when missing pattern are analysed through the notification year. It can be seen that the variables "Clinical_Stage", "Diagnosis_HIV", "Notification_HIV" and "Pregnant_women" have a clear dependence on the notification year, being the information present on the most recent years. The missing rates of the demographical variables "Nationality" and "Residency" is fairly resolvable throughout the years. The missing patterns associated with the variables "Year_Prob_Infection", "Death_date" and "First_symptoms" have consistent missing values throughout all the notification years. Although the variables "CD4", "CD8" and "T4_T8" have high rates of missing

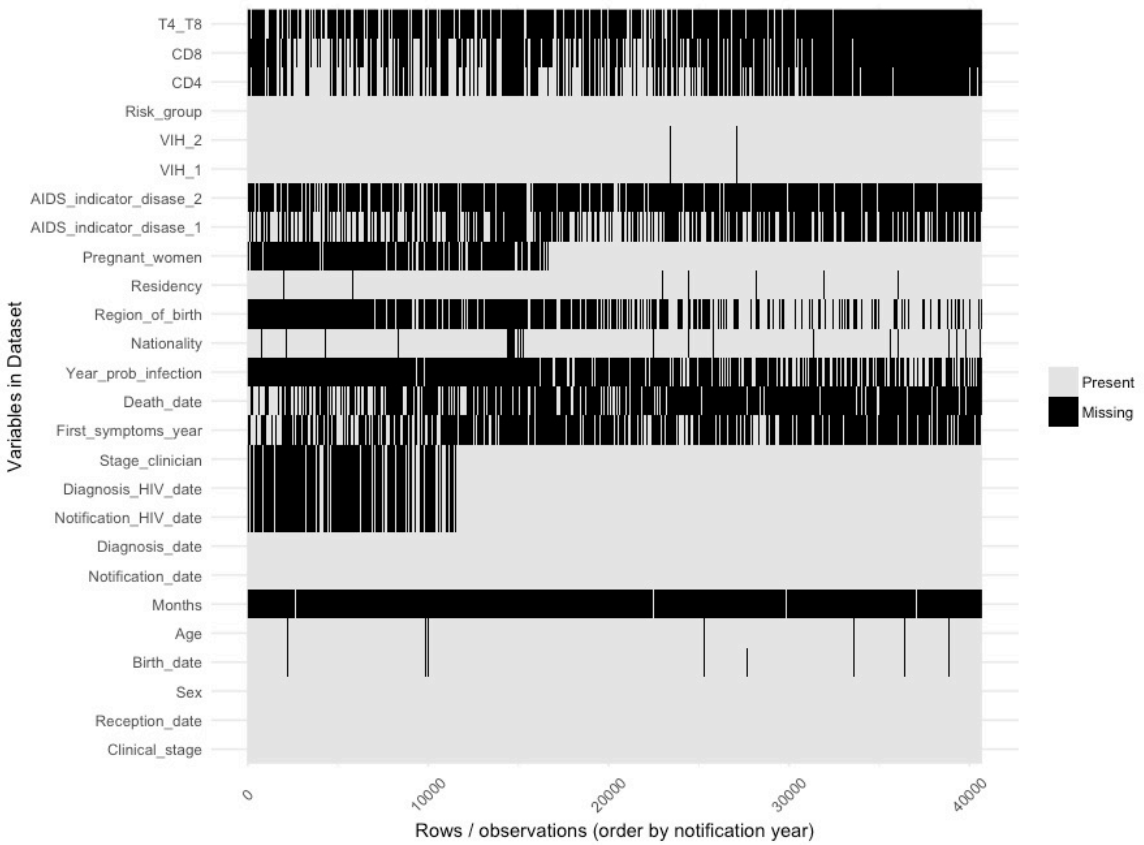


Figure 3.13: Patterns of missing values per notification year

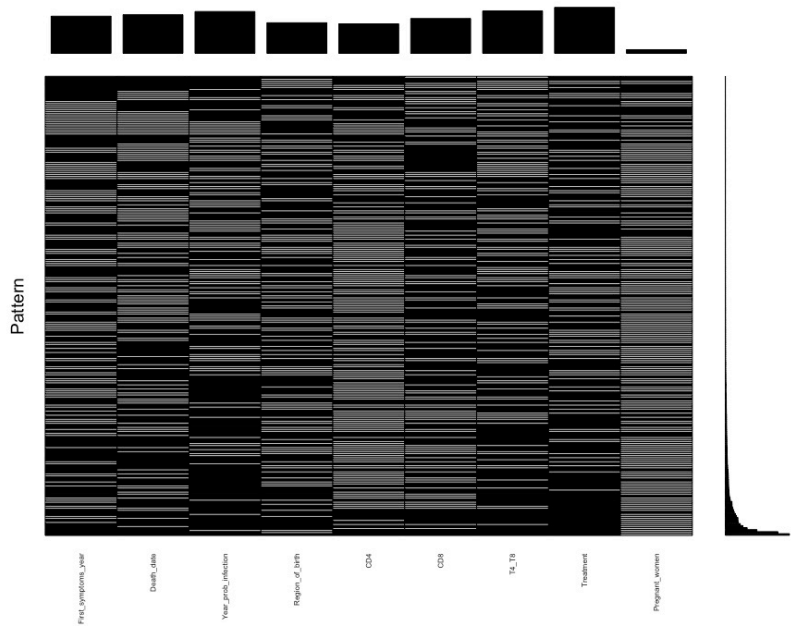
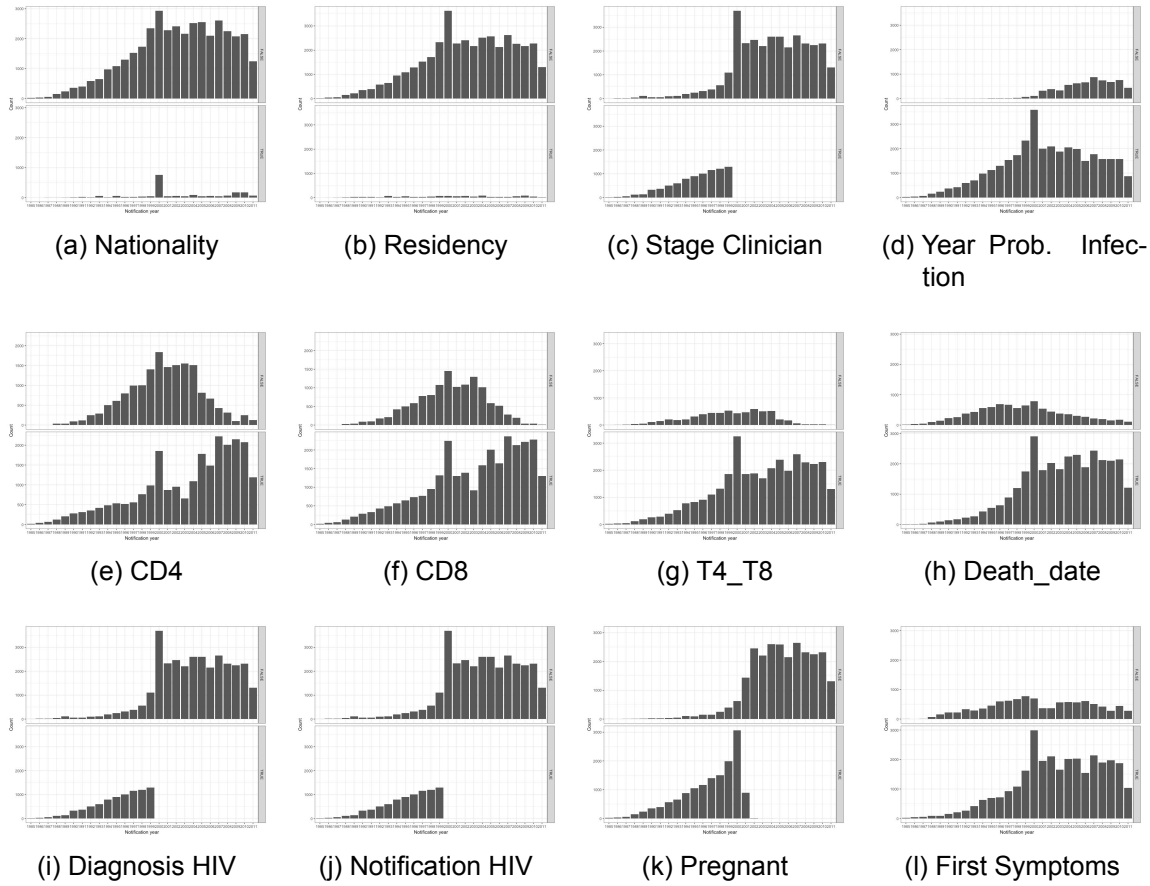


Figure 3.14: Missing values patterns

values throughout the years, that incidence is higher over the last years, since in the 2005 revision these variables were dropped from the notification forms (Figure 3.15.

The number of cases with missing data in some variables are very high (> 5%); nevertheless, it is not reflected on the key variables for the study. Because the rate of missing



(a) Nationality (b) Residency (c) Stage Clinician (d) Year Prob. Infection

(e) CD4 (f) CD8 (g) T4_T8 (h) Death_date

(i) Diagnosis HIV (j) Notification HIV (k) Pregnant (l) First Symptoms

Figure 3.15: Missing values patterns against notification year. The top panel represents the number of observed cases per notification year in each variable while the lower panel represent the number of missing observations per notification year.

values is high, the chosen imputation method will exert a high degree of influence over the results, and differences among competing methods will be magnified. Effects will also be large because of the unusually strong nature of the missingness mechanism [150]. Moreover, there is a clear multilevel nature of this dataset. The observations are nested within health providers which in turn are nested in ARS with administrative autonomy.

Due to the high percentage of missing values and the nested nature of the data, we consider that there is not enough support for uniquely identifying the population each case is originated from, and so we are unable to perform multiple imputations on the variables mentioned above.

At last the timeliness of the data; in these settings, it can be defined as the time between the diagnosis of a case and its proper reporting to the surveillance system. Similarly to all the surveillance systems, reporting delay is an issue present since the beginning (Figure 3.16). This issue will be addressed in chapters 4 and 5.

For augmenting the information of this data set, the names of the Health Providers were matched against the official list of Health Care Providers provided by the ACSS. This list cross-classified each entity with the Health Region Location and type of provided services, such as, hospital care and primary care services.

A crucial feature of a public health surveillance system is the data representativeness and thus if the results are generalizable to the population under surveillance, i.e. have high external validity [162]. When a disease or other outcome of interest is highly prevalent in the general population, and a large proportion of the population comes into at least periodic contact with health services, routine reporting by health clinics and other service-providing

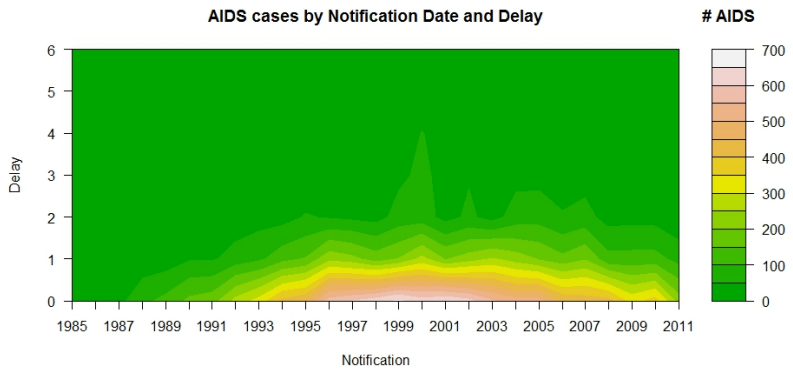


Figure 3.16: Number of AIDS cases per notification year and length of the reporting delay.

institutions suffices as a surveillance mechanism [162].

The entire Portuguese is under HIV - AIDS surveillance but representativeness is unclear. The epidemic is small, concentrated in smouldered high-risk subpopulations whose members are often reluctant to get tested, care or, once diagnosed, often omit important informations. The mapping of these groups relies on easily identifiable of its elements and in some cases it is not possible such in cases of IDU's who have drugs delivered and inject at home, some MSM, clandestine sex workers and their clients... Testing is mandatory for military and strongly recommended for pregnant women, all most-at-risk population such as IDU's and prisoners and HIV status discordant couples. But the high percentage of late diagnosis is an indicator that, individuals do not recognize the transmission modes, symptoms, fear of the result or stigma and discrimination and so just when the severe symptoms emerge the case enters the system. In these settings it is unlikely that the system is capable of producing representative surveillance data. Moreover, under-reporting is a well-recognized problem of the system. It is important to notice that the Portuguese system although classified as comprehensive 'theoretically' covering the all population, it was also classified as passive so the public health agency do not stimulate the reporting nor give feedback about the process.

3.6 Summary

Data cleaning refers to the correction or amelioration of data problems, including missing values, incorrect or out-of-range values, responses that are logically inconsistent with other responses in the database, and duplicate observations. Though we dream with perfection, in reality, 'clean data' is a relative term. Metadata documents are of crucial importance and documents explaining which data elements should be cleaned, with the "description of data validation rules or logical checks for out-of-range values, how missing values and values that are logically inconsistent can be handled, and discussing how duplicate patient observations can be identified and managed" are of great importance. [94].

Data pre-processing and knowledge retrieving, is a multidisciplinary discipline involving topics from statistics, computer science and domain knowledge experts, it is always on data pre-processing and analysis [16].

This chapter describes methods for identifying the main issues in dataset that influence the common dimensions of data quality and provides some correction methods. It proposes two processes: one process for detection of data issues and the another for repairment and improvement of the information in the dataset.

In the case of the Portuguese HIV - AIDS surveillance system, detected barriers were:

illegal values, misspellings, missing data, inconsistencies in disease stage case definition, ambiguous data, outliers, duplicated records and timeliness issues. Most of these barriers were solved with the exception of timeliness, which will be discussed on chapter 4 and on chapter 5.

Chapter 4

Mathematical models for Reporting Delay Estimation

Modelling in science remains, partly at least, an art. Some principles do exist, however, to guide the modeller. A first, though at first sight, not a very helpful principle, is that all models are wrong; some, though, are more useful than others and we should seek those. A second principle is not to fall in love with one model to the exclusion of alternatives. Data will often point with almost equal emphasis at several possible models and it is important that the statistician recognise and accept this.

McCullagh and Nelder [17]

One major component for the control of infectious epidemics is the timely identification of HIV patients and the definition of adequate public health strategies. The number of cases is determined by the collection of diagnosed cases within a health care system. However, typically, a Surveillance System depends on several stakeholders, such as health care providers, the general population and laboratories, which challenges an efficient and effective reporting of diagnosed cases [24]. Delays may be intrinsic of the disease natural history and / or external processes such as time taken to reach a diagnosis or to get the case reported. One issue of particular interest and that often arises is the administrative delay regarding the time mediating from the identification of the HIV-AIDS related event to its national reporting [24]. So, the use of surveillance data requires adjustments of the under-reported case load caused by the delay in reporting within an observation time frame [163].

In this chapter we overview the statistical concepts and fundamentals in section 4.1. This section is divided into two subsections: 4.1.1 where the main statistical and data mining models used for solving the central research questions are presented and discussed, and 4.1.2 discusses some general considerations on model selection and validation. The first presents a brief introduction to: the generalized linear models, the multilevel models, survival analysis, the k-nearest neighbour algorithm, the longitudinal k-means algorithm, naive-bayesian classifier, the multilayer perceptron and support vector machines are briefly described. Section 4.2 reviews the literature, describing strengths and limitations of the main works. Essentially, the traditional approaches can be grouped into two groups: the joint modelling discussed in section 4.2.1 and the separate or partial modelling discussed in section 4.2.2. In section 4.2.3 the data structure is described and section 4.2.4 computes the likelihood for joint models.

4.1 Concepts and Fundamentals

Although simplifying the reality, statistical models investigate the effect of some explanatory variables on some response variables, mimicking the data structure. A well - fitted model has several benefits:

- the structural form of the model describes the patterns of association and interaction;
- estimates of the model parameters determines the strength and importance of the effects;
- inferences about the parameters evaluate which explanatory variables affect the response variable $\mathbf{Y}$, while controlling for possible confounding variables ;
- model's predicted values may smooth the data by providing improved estimates of the mean of $\mathbf{Y}$ at possible explanatory variables values [164].

Modelling in science remains, partly at least, an art [17]. But some principles do exist. Some important steps to be followed include:

1. the understanding of the problem.
2. an adequate experimental planning and data collection.
3. a complete exploratory data analysis, if necessary with data cleaning.
4. model fitting through the following stages:
   (a) Postulate the model.
   (b) Estimate parameters.
   (c) Select and validate the model.
   (d) Return to 4a if model does not satisfy the criterion defined in 4c.
   (e) Use model to investigate the effects of some explanatory variables on response variable(s).

Understanding the full scope of the problem is essential to the model purpose and goals, identification of the important variables, and definition of their scale and relationships.

Data is a critical element in any modelling process. The way it is collected may determine the significant power and conclusions of posterior statistical inferences, thus turning data cleaning into a crucial initial process. Also, it may highlight data quality issues (for example, the values plausibility, as discussed in Chapter 3) [165].

Steps 4b to 4e are commonly named Model Fitting and are the kernel of any modelling-building process. The goal is to develop a functional relationship between random variables that accurately describes the intrinsic variability structure with low bias. The observed data are supposed to be a realization of these random variables or, of a measurable function of these random variables [166].

In step 4a, one must specify a model class and select a model structure, namely the set of covariates or explanatory variables, the equation linking the response to the explanatory variables and the reasonable probability distribution of the error structure in the parametric models.

Parameter estimation varies according to the model at hand but maximum likelihood and least squares methods are among the most used methods in parametric modelling. If conditions for parameter estimation are imposed, then validation of those conditions has to be performed. The model validation is usually the next step, evaluating the ”closeness” between the predicted and the true values of the response for given values of the explanatory variables. [165].

In this section are presented some models ranging from the generalized linear models to some data mining approaches.

When postulating a model we must take into consideration the distribution of the response, the selection of relevant co-variable and the choice of the functional relationship between the expected value and the linear predictor. The Generalized Linear Models (GLMs) form a unified approach of regression models that extend the linear regression model to a large set of non-Gaussian responses and to non-identity relationships between the mean response and the linear predictor [167, 168, 169]. The postulation of the model will be focused in section 4.1.1.

Random-effect models, and more generally multilevel models, are essentially an expansion of the generalized linear models to correlated and nested data. The modelling of the inter-individual variation makes use of the so-called random-effects, which are parameters with a structure of random variables. They are particularly appropriate for research designs where data for participants are organized at more than one level [170].

Survival analysis is a sub-field of statistics that analyses and models data where the outcome is time until the occurrence of an event of interest [171]. This framework presents many challenges, including the incorporation of information from individuals not experiencing any event during the monitoring period - censored observations. Due to censoring, the commonly used statistical methods do not apply straightforwardly. Statistical approaches have been widely developed in the literature to overcome that threat also, many machine learning algorithms were adapted to effectively address this characteristic [171].

The term “data mining” refers to a collection of techniques that provide the necessary actions to retrieve and gather knowledge from an exhaustive assemblage of data and facts [172]. In particular, they can uncover new biomedical and health care knowledge for clinical and administrative decision-making as well as generate scientific hypotheses from large experimental data, clinical databases, and / or biomedical literature [173]. Data mining models can be classified into two categories: descriptive (or unsupervised learning) and predictive (or supervised learning) [174]. Descriptive data mining consists of a collection of techniques aiming to discover unknown patterns or relationships in data. This exploratory analysis includes clustering, association, and sequence discovery [174]. Predictive data mining infers prediction rules from data. It includes tasks such as classification, regression, time series analysis, and prediction [173]. Classification is the most frequently used data mining method with a predominance of the implementation of k-Nearest Neighbour (k-NN) algorithm, k-means algorithm, Bayesian classifiers, Neural Networks, and SVMs [175].

Having selected the class of the model, it is necessary to choose from the set of all possible parameters values the ones that make the fitted values closer to the observed ones. So, a measure of goodness-of-fit must be defined. The optimal parameters will be those minimizing a certain ”closeness” criterion. One may use information criteria, such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC), for model selection among a finite set of models, once the likelihood function is defined for those models; or evaluate the fitting by efficiently data reuse through cross-validation or bootstrap. A good model for a specific problem must achieve the equilibrium of three characteristics: must be adequate, be minimal and easily interpreted [167].

4.1.1 Statistical and Data Mining Models

Generalized Linear Model

Terminology A data set will be represented by a matrix, a two-dimensional array in which the rows are indexed by experimental or survey units and the columns correspond to the variables such as measurements.

The variables regarded as responses or dependent variables are the variables whose values are believed to be affected by the covariates variables [17].

All variables may be quantitative or qualitative. The first type of variables take on numerical values while qualitative, some times referred to as *factors*, take non-numerical values or *levels* from a finite set of values or labels.

Let $\mathbf{Y} = [Y_1, \dots Y_n]^T$ be a random variable. The set of covariates, which it is believed to explain some of the variability of $\mathbf{Y}$, is arranged as an $n \times p$ matrix $\mathbf{X} = [X_1, \dots, X_p]$; here, each row of $\mathbf{X}$ refers to a different observation, and each column to a different covariate, which may be deterministic or random variables. Associated with each covariate is a coefficient or parameter, usually unknown. Each $\mathbf{X_i}, i = 1, \dots, p$ and $\mathbf{Y}$ may be continuous, discrete, polytomous or even binary. When the $\mathbf{X_i}$ is qualitative with k categories, they must be coded into $k-1$ dummy variables.

In a linear regression context, the relation between $\mathbf{Y}$ and $\mathbf{X}$ can be written as

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{4.1}$$

where $\mathbf{Z}$ is a matrix of dimension $n \times (p+1)$, with first column consisting of 1's and the rest given by $\mathbf{X}$, $\boldsymbol{\beta} = [\beta_0, \dots, \beta_p]^T$ of dimension $p+1$ and ε is a vector of random errors of dimension $n \times 1$. In matrix form, equation 4.1 become:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \tag{4.2}$$

which is equivalent to

$$\mathbf{Y}|\mathbf{X} \sim N\left(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}\right) \tag{4.3}$$

in particular, $E(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$.

The GLMs are composed by three parts: a random component, a systematic component and a link function relating the two components [176]:

- **The random component, response distribution or 'error structure'**:

  Given $\mathbf{X} = [X_1, \dots, X_p]$, the $Y_i\,(i = 1, \dots, n)$ are (conditionally) independent random variables with means $E\left(Y_i|\mathbf{X}_i\right) = \mu_i$ and all following the same distribution belonging to the exponential family (to be defined below) .

- **The systematic or structural component** :

  Consists of a specification of the vector $\boldsymbol{\mu_i}$ in terms of unknown parameters $\beta_0, \beta_1, \dots, \beta_p$. Most of the times, a linear predictor $\boldsymbol{\eta} = \mathbf{Z}\boldsymbol{\beta}$ is used.

- **The link function between the random and systematic component**: If a linear predictor $\boldsymbol{\eta}_i = \mathbf{Z}_i\boldsymbol{\beta}_i$ is used, the conditional expected value $\boldsymbol{\mu}_i$ is related to the linear predictor by a monotonic differentiable function h such that:

  $$\boldsymbol{\mu}_i = h\left(\boldsymbol{\eta}_i\right) = h\left(\mathbf{Z}_i\boldsymbol{\beta}_i\right).$$

  This is the component that describes how the location of the response distribution changes with the exploratory variables.

  The function $g = h^{-1}$ is called the link function. So

  $$g\left(\boldsymbol{\mu}_i\right) = \boldsymbol{\eta}_i = \mathbf{Z}_i\boldsymbol{\beta}_i = \beta_{0i} + \beta_{1i}X_{1i} + \ldots + \beta_{pi}X_{pi},$$

describes the relationship between the mean of the response and the linear predictor.

The choice of this link function depends on the problem and on the distribution of $\mathbf{Y}$.

For clarity, the index i will be omitted whenever the comprehension is not compromised.

Classical regression models assume a normal distribution on the random component and the identity function such that the linear predictor $\boldsymbol{\eta}$ is equal to $\boldsymbol{\mu}$.

GLMs use a special family of distribution functions enjoying nice statistical properties, the exponential family [176]. A brief overview of this family is described below .

Exponential family distributions

Definition 4 *A random sample Y has a distribution function belonging to the exponential family if its density function can be written in the form*

$$f_Y(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}, \tag{4.4}$$

for some specific real functions $a(.)$, $b(.)$ and $c(.,.)$ and real values ϕ and θ.

The parameters, θ is the canonical form of the location parameter, while ϕ is the dispersion parameter which, in general is known [167, 17, 176]. Different specifications for the functions $a(.)$ and $b(.)$ lead to different distributions. For some distributions (e.g. binomial), ϕ is equal to 1 thus not requiring any estimation. The cumulative function is $b(\theta)$ and $b''(\theta)$ is the variance function.

It can be shown ([17]) that the mean and variance of a random variable $\mathbf{Y}$ following a distribution from the exponential family are given by, respectively:

$$\begin{aligned} E(Y) &= \mu = b'(\theta) \\ V(Y) &= b''(\theta)\, a(\phi) \end{aligned}$$

The most important distributions of the exponential family are the Binomial distribution, used when $\mathbf{Y}$ takes on dichotomous values; the Poisson for count data, the Normal distribution and the Gamma and Inverse Gaussian distributions when the response consists of non-negative continuous values.

Link function The simplest link function is the *identity link* used in the linear regression model.

Other link functions allow $\boldsymbol{\mu}$ to be non-linearly related to the linear predictor being the most popular the ones described in Table 4.1. The choice of the link function should be compatible with the distribution of the response and must take into consideration the problem, the observed data and the interpretation of the results.

The canonical link function connects μ, η and θ such that $\eta \equiv \theta$, then this link is called canonical.

Table 4.1: Common canonical link functions [167]

Family	Canonical Link	Variance Function
Normal	$\eta = \mu$	1
Poisson	$\eta = log\mu$	μ
Binomial	$\eta = log\left(\frac{\mu}{1-\mu}\right)$	$\mu(1-\mu)$
Gamma	$\eta = \mu^{-1}$	μ^2
Inverse Guassian	$\eta = \mu^{-2}$	μ^3

Some characteristics, such as support (range of Y), parameters, cumulative functions

and variances functions of these distributions are summarized in Table 4.2.

Table 4.2: Characteristics of some common univariate distributions in the exponential family. Based in [167].

	Normal	Binomial	Poisson	Gamma	Inverse Gaussian
Notation	$N\left(\mu,\sigma^2\right)$	$B\left(m,\pi\right)/m$	$P\left(\mu\right)$	$G\left(\upsilon,\frac{\upsilon}{\mu}\right)$	$IG\left(\mu,\sigma^2\right)$
Support	$(-\infty,\infty)$	$\left\{0,\frac{1}{m},\ldots,1\right\}$	$\{0,1,\ldots\}$	$(0,\infty)$	$(0,\infty)$
θ	μ	$\ln\left(\frac{\pi}{1-\pi}\right)$	$\ln\left(\lambda\right)$	$-\frac{1}{\mu}$	$-\frac{1}{2\mu^2}$
$a\left(\phi\right)$	σ^2	$\frac{1}{m}$	1	$\frac{1}{\upsilon}$	σ^2
ϕ	σ^2	1	1	$\frac{1}{\upsilon}$	σ^2
$c\left(y,\phi\right)$	$-\frac{1}{2}\left(\frac{y^2}{\phi}+\ln\left(2\pi\phi\right)\right)$	$\ln\binom{m}{my}$	$-\ln y!$	$\upsilon\ln\upsilon-\ln\Gamma\left(\upsilon\right)+\left(\upsilon-1\right)\ln y$	$-\frac{1}{2}\left(\ln\left(2\pi\phi y^3\right)+\frac{1}{y\phi}\right)$
$b\left(\theta\right)$	$\frac{\theta^2}{2}$	$\ln\left(1+\exp(\theta)\right)$	$\exp(\theta)$	$-\ln\left(-\theta\right)$	$-\left(-2\theta\right)^{1/2}$
$b'\left(\theta\right)=E\left(Y\vert\theta\right)=\mu\left(\theta\right)$	θ	$\pi=\frac{\exp(\theta)}{1+\exp(\theta)}$	$\lambda=\exp\left(\theta\right)$	$\mu=-\frac{1}{\theta}$	$\mu=\left(-2\theta\right)^{1/2}$
$b''\left(\theta\right)=V\left(\mu\right)$	1	$\pi\left(1-\pi\right)$	λ	μ^2	μ^3
$var\left(Y\right)$	σ^2	$\frac{\pi(1-\pi)}{m}$	λ	$\frac{\mu^2}{\upsilon}$	$\mu^3\sigma^2$

Maximum Likelihood Estimation Although, in some special cases, explicit mathematical expressions for the maximum likelihood estimator can be obtained, numerical methods are usually needed. Typically these methods are a form of iteratively re-weighted least squares and are based on the Newton-Raphson algorithm [165].

Consider independent and identically distributed random variables $Y_1,\ldots,Y_n$. In a GLM, one wishes to estimate the parameters $\boldsymbol{\beta}$, usually denoted by regression coefficients or parameters. The standard way of doing so is by maximum likelihood. In a GLM context, the likelihood function is given by

$$L(\theta,\phi;y,x)=\prod_{i=1}^{n}f_Y\left(y_i;\theta_i,\phi\right). \tag{4.5}$$

For maximization purposes, it is better to consider the logarithm of the likelihood function, thus for independent observation it is considered maximizing

$$l(\theta,\phi;y,x)=\log L(\theta,\phi;y,x)=\sum_{i=1}^{n}\frac{y_i\theta_i-b\left(\theta_i\right)}{a\left(\phi\right)}+c\left(y_i;\phi\right). \tag{4.6}$$

Since $\mu=b'\left(\theta\right)$, $g\left(\mu\right)=\eta$ and $\eta=\beta_0+\beta_1X_1+\ldots+\beta_pX_p$, the likelihood in (4.5) can be represented by $L(\boldsymbol{\beta};y,x)$ and the log-likelihood (4.6) by $l(\boldsymbol{\beta};y,x)$.

Inference The two main tools of statistical inference are confidence intervals and hypothesis tests [165]. Confidence intervals are increasingly regarded as more useful than hypothesis tests because their width provides a measure of the precision with which inferences can be made [165].

The hypothesis tests are performed by comparing nested models where the parameters of the model with the fewest parameters (M_0) is a subset of the parameters of the other model (M_1). In the case of generalized linear models, these two models must have the same response distribution and the same link function [165].

The two common approaches for model comparison are the log-likelihood ratio test and the Wald test (usually applied to test for the significance of a regression parameter).

In general, let $\boldsymbol{\beta}_{M_1}$ denote the p parameters of the model M_1 and $\hat{\beta}_{M_1}$ be the maximum likelihood estimator of $\boldsymbol{\beta}_{M_1}$. In the same way, let $\boldsymbol{\beta}_{M_0}$ denote the r (with $p>r>0$) parameters of the model M_0 and $\hat{\beta}_{M_0}$ be the maximum likelihood estimator of $\boldsymbol{\beta}_{M_0}$ Let the log-likelihood function for the M_1 model evaluated at $\hat{\beta}_{M_1}$ be $l(\hat{\beta}_{M_1};y,x)$ and $l(\hat{\beta}_{M_0};y,x)$ denote the maximum value of the log-likelihood function for the model M_0. Then the log likelihood-ratio statistics is

$$G = -2\log\left(\frac{L(\hat{\beta}_{M_0}; y, x)}{L(\hat{\beta}_{M_1}; y, x)}\right) = -2\left(l(\hat{\beta}_{M_0}; y, x) - l(\hat{\beta}_{M_1}; y, x)\right). \tag{4.7}$$

Large values of G suggest that the model M_0 is a poor descriptor of the data relative to M_1. The critical region of G is determined throughout its sampling distribution. Large sample theory states that the asymptotic distribution of likelihood ratio statics under the usual regularity conditions and under H_0 is χ^2_{p-r}:

$$G \overset{a}{\sim} \chi^2_{(p-r)}. \tag{4.8}$$

The Wald test is a quadratic approximation of G. It is a simple test that uses the sampling distribution of the parameter maximum likelihood estimate. If the response variable is Normally distributed, the sampling distribution used for inference can often be determined exactly. For other distributions the large-sample asymptotic results based on the Central Limit Theorem can be used. Under the null hypothesis

$$\begin{aligned} H_0 &: \beta = 0 \\ H_1 &: \beta \neq 0 \end{aligned} \tag{4.9}$$

$$W = \left(\frac{\hat{\beta}}{\hat{se}\left(\hat{\beta}\right)}\right)^2 \overset{a}{\sim} \chi^2_{(1)} \tag{4.10}$$

or, equivalently

$$W = \frac{\hat{\beta}}{\hat{se}\left(\hat{\beta}\right)} \overset{a}{\sim} N(0,1). \tag{4.11}$$

Again, high values of W suggest that the variable associated to the the parameter β is significant.

The goodness-of-fit test is a model comparison test where M_0 is identified as the model of interest and the M_1 is the full or saturated model were the data is explained exactly, typically needing n parameters for n data points [168]. Also, the full model gives a measure of how well any model could possibly fit and so we might consider the difference between the log-likelihood for the full model, $l\left(\tilde{\theta}, \phi; y, x\right)$ and that for the model under consideration, $l\left(\hat{\theta}, \phi; y, x\right)$,expressed as a likelihood ratio statistic:

$$2\left(l\left(\tilde{\theta}, \phi; y, x\right) - l\left(\hat{\theta}, \phi; y, x\right)\right). \tag{4.12}$$

Provided that the observations are independent and for an exponential family distribution, when $a_i(\phi) = \frac{\phi}{w_i}$ with w_i this a known weight simplifies to:

$$\sum_i \frac{2w_i\left(y_i\left(\tilde{\theta}_i - \hat{\theta}_i\right) - b\left(\tilde{\theta}_i\right) + b\left(\hat{\theta}_i\right)\right)}{\phi} \tag{4.13}$$

which can simply be written as $d\left(\hat{\theta}; y\right)/\phi$ where $d\left(\hat{\theta}; y\right)$ is called the deviance and $d\left(\hat{\theta}; y\right)/\phi$ is called the scaled deviance.

Model Checking As with standard linear models, it is important to check the adequacy of the assumptions supporting GLMs. The diagnostic methods for GLMs mirror those used for the Gaussian linear models but some adaptations are necessary and, depending on the type of the response, not all diagnostic methods will be applicable [168].

The choice of the response distribution is done by scientific insight about the data or by comparing the model fit for a variety of different distributions selected based on prior assumptions [177].

The diagnostic methods can be divided into two groups: methods to detect single cases or small groups of cases that do not fit the pattern of the rest of the data and methods to check the assumptions of the model. The latter, which is the focus here, can be divided into those that check the structural form of the model, such as the choice and transformation of the predictors, and those that check the random component of the model[168].

The following is based on the book of Faraway on GLMs. He suggests that the diagnostic methods should be performed as follows ([168]) :

- checks of the random component of the model: plot of Pearson residuals versus fitted values and versus the linear predictor $\hat{\eta}$. Where Pearson residuals are defined as:

$$r_P = \frac{y - \hat{\mu}}{\sqrt{V(\hat{\mu})}} \tag{4.14}$$

  where $V(\mu) = b''(\theta)$.

  - Check for the presence of nonlinear relationships between the predicted values and the residuals. If so, this would indicate a lack of fit and that a change in the model is needed. It is recommended to make changes only on the predictors since that involves the least disruption to the GLMs.
  - Check the variance of the residuals with respect to the fitted values. The GLMs assumptions require constant variance in that plot and a violation of such condition implies a change in the model. For all GLMs but Gaussian, the variance function is non-constant; however, if the deviance residuals are used, the variation is scaled out and a constant variance is expected in such a plot. The deviance residuals can be defined as:

$$r_D = sign(y - \hat{\mu}) \sqrt{d_i} \tag{4.15}$$

    where d_i is the individual deviance contribution.

  Note that, in some cases, plots of the residuals may not be particularly helpful. In the case of discrete responses, such as binomial responses with small group sizes and Poisson responses with small values, the residuals can only take few distinct values. Moreover, the plots tend to show curved lines consisting of points corresponding to the limited number of observed values. Such artefacts can obscure the main purpose of the plot. More difficulties arise if the covariates are measured in a very different scales.

- checks for the structural component of the model - plots of the response versus each covariate. These plots allow the investigation of the nature of the relationship between the predictors and the response. The interpretation must be done carefully since these simple representations do not take into account the effect of other predictors.
- check the nature of the relationship between the predictors and the response plotting the response against the predictors.
- In any scientific context, predicted and observed values should also be confronted [169].

Multilevel Models

Many common statistical models can be expressed as linear models that incorporate both *fixed effects*, which are parameters associated with an entire population or with certain repeatable levels of experimental factors, and *random effects*, which are associated with individual experimental units drawn at random from a population [178]. A model with both *fixed effects* and *random effects* is called a mixed-effects model, hierarchical model or even a multilevel model.

The primary goal of the multilevel model is to describe relationships between a response variable and some covariates in data that are grouped according to one or more classification factors [178].

Definition 5 *A fixed effect is an unknown constant that we try to estimate from the data. A random effect is a random variable hence it requires the estimation of the parameters of its distribution.*

As extensions of GLMs, the linear predictor in multilevel models is linear in the fixed and random-effects.

In this section we follow Pinheiro and Bates formulation of multilevel models described in [178] and [179].

Lets consider a single level of random-effects. The linear mixed-effects model with response $\mathbf{Y}_i \in \mathbb{R}^{n_i}$ for the ith group is:

$$\begin{aligned} \mathbf{Y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \qquad i = 1, \ldots, M, \\ \mathbf{b}_i &\sim N\left(0, \Sigma_\theta\right), \boldsymbol{\epsilon}_i \sim N\left(0, \sigma^2\mathbf{I}\right), \end{aligned} \tag{4.16}$$

where $\boldsymbol{\beta}$ is the p - dimensional vector of fixed effects, $\mathbf{b}_i$ is the q-dimensional vector of random effects, $\mathbf{X}_i$ (of size $n_i \times p$) and $\mathbf{Z}_i$ (of size $n_i \times q, q < p$) are known fixed-effects and random-effects regressor matrices, and $\boldsymbol{\epsilon}_i$ is the n_i-dimensional within-group error vector with a Gaussian distribution. The random effects $\mathbf{b}_i$ and the within-group errors $\boldsymbol{\epsilon}_i$ are assumed to be independent for different groups and to be independent of each other for the same group [178].

The distribution of the random effects vectors $\mathbf{b}_i$ is completely characterized by its variance–covariance matrix Σ_θ which must be symmetric and positive - definite. The variance-covariance matrix can be written in a more convenient form in terms of a relative covariance factor, Δ_θ, which is a $q \times q$ matrix, depending on a variance-component parameter, θ such as:

$$\Sigma_\theta = \sigma^2\Delta_\theta\Delta_\theta^T. \tag{4.17}$$

If Σ_θ is positive-definite then Δ_θ exists although it is not unique. Given this decomposition, $\mathbf{b}$ can be written in terms of a spherical random variable $\mathbf{U} \sim N\left(0, \sigma^2\mathbf{I}_q\right)$ such as

$$\mathbf{b} = \Delta_\theta\mathbf{U}. \tag{4.18}$$

For this model, the penalized residual sum of squares is the sum of the residual sum of squares, measuring fidelity of the model to the data, and a penalty on the size of $\mathbf{U}$, measuring the complexity of the model [179]:

$$r^2\left(\theta, \boldsymbol{\beta}, \mathbf{U}\right) = \parallel \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\Delta_\theta\mathbf{U} \parallel^2 + \parallel \mathbf{U} \parallel^2 \tag{4.19}$$

For minimizing r^2 with respect to $\mathbf{U}$, that is

$$r^2_{\boldsymbol{\beta},\theta} = \min_{\mathbf{U}} \left\{ \parallel \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\Delta_\theta\mathbf{U} \parallel^2 + \parallel \mathbf{U} \parallel^2 \right\} \tag{4.20}$$

it is necessary to determine the sparse Cholesky factor, $\mathbf{L}_\theta$, which is a lower triangular $q \times q$ matrix satisfying

$$\mathbf{L}_\theta \mathbf{L}_\theta^T = \Delta_\theta^T \mathbf{Z}^T \mathbf{Z} \Delta_\theta + \mathbf{I}_q \tag{4.21}$$

where $\mathbf{I}_q$ is the $q \times q$ identity matrix.

The estimate of $\boldsymbol{\beta}$ is given by the minimum of the function

$$r_\theta^2 = \min_{\mathbf{u},\beta} \left\{ \| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\Delta_\theta \mathbf{U} \|^2 + \| \mathbf{U} \|^2 \right\} \tag{4.22}$$

and is called the conditional estimate of $\boldsymbol{\beta}$ given θ and represented as $\hat{\boldsymbol{\beta}}_\theta$.

The deviance of the model, given the data, $\mathbf{y}$, is [179]

$$d\left(\theta, \boldsymbol{\beta}, \sigma; \mathbf{y}\right) = n \log\left(2\pi\sigma^2\right) + \log\left(|\mathbf{L}_\theta|^2\right) + \frac{r_{\beta,\theta}^2}{\sigma^2} \tag{4.23}$$

where $|\mathbf{L}_\theta|$ denotes the determinant of $\mathbf{L}_\theta$.

Minimizing r^2 in terms of $\boldsymbol{\beta}$ and $\mathbf{u}$ gives the so called conditional mode of the spherical random effects and the conditional estimate, $\hat{\beta}_\theta$, of the fixed effects. At the conditional estimate of the fixed effects $\hat{\boldsymbol{\beta}}_\theta$ the deviance, can be written as

$$d\left(\theta, \boldsymbol{\beta}, \sigma; \mathbf{y}\right) = n \log\left(2\pi\sigma^2\right) + \log\left(|\mathbf{L}_\theta|^2\right) + \frac{r_\theta^2}{\sigma^2} \tag{4.24}$$

whose minimum with respect to σ^2 is the conditional estimate

$$\hat{\sigma}_\theta^2 = \frac{r_\theta^2}{n} \tag{4.25}$$

which provides the profile deviance:

$$\tilde{d}\left(\theta; \mathbf{y}\right) = \tilde{d}\left(\theta, \hat{\boldsymbol{\beta}}_\theta, \hat{\sigma}_\theta; \mathbf{y}\right) = \log\left(|\mathbf{L}_\theta|^2\right) + n\left[1 + \log\left(\frac{2\pi r_\theta^2}{n}\right)\right], \tag{4.26}$$

a function of just θ.

The maximum likelihood estimate of θ is the value that minimizes the profiled deviance in equation (4.26) [179]. The minimum of this deviance is obtained throughout numerical optimization, such as the EM algorithm or Newton - Raphson algorithms.

The elements of the conditional mode of $\mathbf{b}$, also called the best linear unbiased predictors of the random effects, evaluated at the parameter estimates are

$$\tilde{\mathbf{b}}_{\hat{\theta}} = \Delta_{\hat{\theta}} \tilde{\mathbf{u}}_{\hat{\theta}}. \tag{4.27}$$

The restricted or residual maximum likelihood estimates (REML) of the variance components are often preferred over the maximum likelihood estimates since they have a lower bias than the latter. [179]. This approach is a particular form of maximum likelihood estimation that uses a likelihood function calculated from a transformed set of data so that nuisance parameters have no effect [180].

Using REML, the deviance can be can be expressed as

$$d_R\left(\theta, \sigma; \mathbf{y}\right) = -2 \log \int_{\mathbb{R}^p} L\left(\theta, \beta, \sigma; \mathbf{y}\right) d\beta. \tag{4.28}$$

In this setting the penalized residual sum of squares is

$$r_{\theta,\beta}^2 = r_\theta^2 + \| \mathbf{R}_X \left(\beta - \hat{\beta}_\theta\right) \|^2 . \tag{4.29}$$

where $\mathbf{R_X}$ is an upper triangular $p \times p$ matrix satisfying

$$\mathbf{L}_\theta \mathbf{R}_{ZX} = \mathbf{P}\Delta_\theta^T \mathbf{ZX} \tag{4.30}$$

and

$$\mathbf{R}_\mathbf{X}^T \mathbf{R_X} = \mathbf{X}^T\mathbf{X} - \mathbf{R}_\mathbf{ZX}^T \mathbf{R_{ZX}} \tag{4.31}$$

with $\mathbf{P}$ representing a $q \times q$ permutation matrix consisting of permuted columns of the identity matrix, $\mathbf{I}_q$. This matrix aids the reduction of the number of non-zeros in the factor, $\mathbf{L}_q$.

So, the deviance can be re-written as

$$d_R\left(\theta, \sigma; \mathbf{y}\right) = (n-p)\log\left(2\pi\sigma^2\right) + 2\log\left(\mid \mathbf{L}_\theta \mid\mid \mathbf{R_X} \mid\right) + \frac{r_\theta^2}{\sigma^2}. \tag{4.32}$$

and the profile deviance is

$$\tilde{d}_R\left(\theta; \mathbf{y}\right) = 2\log\left(\mid \mathbf{L}_\theta \mid\mid \mathbf{R_X} \mid\right) + (n-p)\left[1 + \log\left(\frac{2\pi r_\theta^2}{n-p}\right)\right]. \tag{4.33}$$

The REML estimates are

$$\hat{\theta}_R = \arg min_\theta \tilde{d}_R\left(\theta; \mathbf{Y}\right) \tag{4.34}$$

and

$$\hat{\sigma}_R^2 = \frac{r_{\hat{\theta}_R}^2}{n-p}. \tag{4.35}$$

The estimate of $\boldsymbol{\beta}$ is commonly taken to be $\hat{\beta}_R = \hat{\beta}_R\left(\theta\right)$.

The models are compared according to the change in the deviance, which is the likelihood ratio test statistic [179]. Bates *et all applied* a signed square root transformation to this statistic and plotted the resulting function, called the profile zeta plot and represented by ζ , versus the parameter value [179]. A ζ value can be compared to the quantiles of the standard normal distribution, $\zeta \sim N(0,1)$ and when the plot exhibits linearity for a given parameter it implies that the likelihood profile is quadratic and thus that Wald approximations would be reasonably accurate [179].

The single-level multilevel model can be extended to accommodate multiple, nested levels of random effects in a similar way.

Survival Analysis

Terminology An important type of data is the time from a well-defined starting point until some event occurs . Data on times until the event, or more commonly, 'duration of survival' or 'survival times', have two important features:

- the times are non-negative and typically have skewed distributions with long tails;
- some of the subjects may survive beyond the study periods so that their actual survival times may not be known; in this case, and other cases where the failure times are not known completely, the data are said to be censored [165].

Random censoring can be broadly categorized into three groups based on when the record of time stops:

- right-censoring, occurs when a subject leaves the study before an event occurs, or the study ends before the event has occurred. So, the observed survival time is less than or equal to the true survival time;

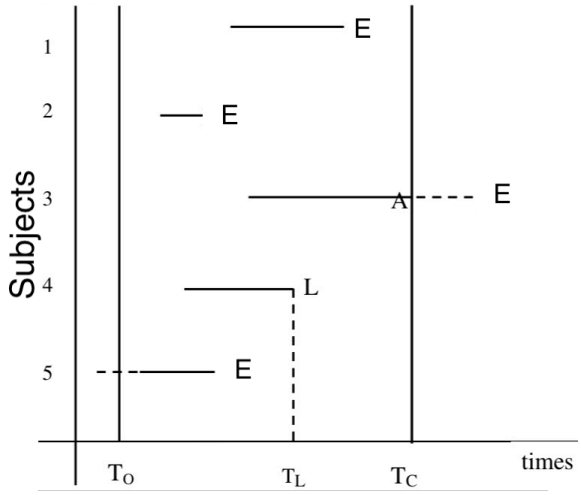


Figure 4.1: Diagram of types of censoring for survival times [171].

- left-censoring, occurs when the event of interest has already occurred before enrolment. For which the observed survival time is greater than or equal to the true survival time;
- interval censoring, the combination of the previous two, for which we only know that the event occurs during a given time interval [171].

It should be noted that the true event occurrence time is unknown in all the three cases.

Other important mechanism of missing information is truncation which is deliberate and due to systematic selection process inherent to the study design. Similar to censoring, it can be right, left or interval.

Note that censoring is a problem different from truncation, because with censored data the researcher has at least some partial information on the censored lifetimes. Right truncation happens when the selection process of the study stops at some point. For example, if a study deals with people who got infected within a specific time period, there is a lack of observation for infections past this specific time period. This lack of observation is considered to be right-truncated.

In Figure 4.1, an illustrative example is given for a better understanding of the definition of censoring and the structure of survival data. Five instances were observed in this study that occurred between T_O and T_C which are the beginning and end of the study period, respectively. E denotes 'event' and A denotes 'alive at the end of the study'; L indicates that the subject was known to be alive at the time shown but then became lost to the study and so the subsequent life course is unknown [165].

For subjects 1 and 2, the entire survival period occurred within the study period. For subject 3, 4 and 5 the survival times were censored and so, are partially observed. More specifically, subject 3 is censored since there was no event occurred during the study period, 4 is censored due to the withdraw or being lost to follow-up within the study time period and 5 since the survival time started before the initial of observation study.

The censoring problem can be stated as:

For a given instance i, represented by a triplet (X_i, y_i, δ_i), where $X_i \in \mathbb{R}^{1\times P}$ is a covariate (feature) vector; δ_i is the binary event indicator, i.e., $\delta_i = 1$ for an uncensored instance and $\delta_i = 0$ for a censored instance; and y_i denotes the observed time and is equal to the survival time T_i for an uncensored instance and C_i for a censored instance i.e.,

$$y_i = \begin{cases} T_i & \text{if } \delta_i = 1 \\ C_i & \text{if } \delta_i = 0. \end{cases} \tag{4.36}$$

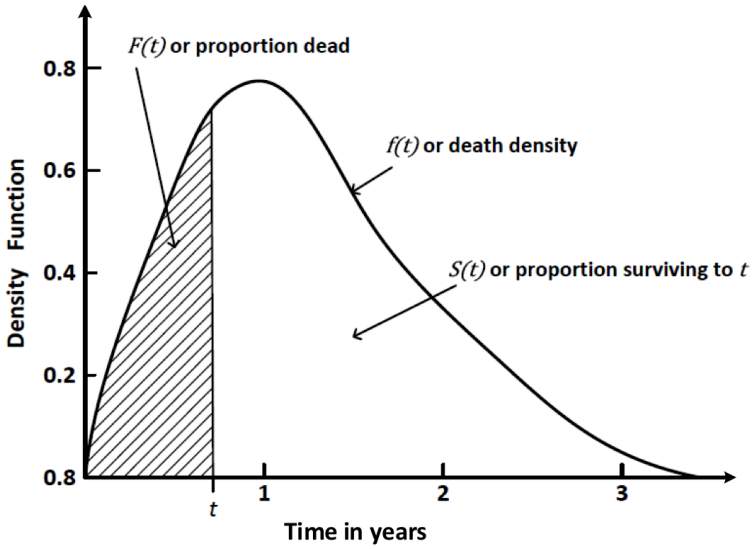


Figure 4.2: Relationship among different entities $f(t)$, $F(t)$ and $S(t)$ [171].

It should be noted that T_i is a partly observed survival time.

The goal of survival analysis is to estimate the time to event of interest T_j for a new instance j with covariates denoted by X_j.

The Survival and Hazard Functions - $S(t)$ and $\lambda(t)$ Let the random variable T denote the survival time and let $f(t)$ denote its probability density functions (also denoted throughout by death function). Then the probability of an individual surviving until at least time t is

$$S(t) = Pr(T \geq t) \tag{4.37}$$

where t is a time of interest. This function S is denoted by survival function.

The survival function monotonically decreases with t, and at $t = 0$ is equal to 1, representing the fact that, at the beginning of the study period, 100% of the subjects 'survive'; or in other words, none of the events of interest has occurred [171].

On the contrary, the cumulative death distribution function, or proportion of death,

$$F(t) = Pr(T \leq t) = \int_0^t f(x)\,dx = 1 - S(t) \tag{4.38}$$

represents the probability of event occurring earlier that t. The death density function $f(t)$ can be obtained by $f(t) = \frac{d}{dt}F(t)$ in the continuous case and by $f(t) = \frac{F(t+\triangle t)-F(t)}{\triangle t}$, where $\triangle t$ represents a small variation in time interval, in the discrete case. Figure 4.2 represents the relationship between these functions.

The **hazard function** $\lambda(t)$, also called the *force of mortality*, the *instantaneous death rate* or the *conditional failure rate* is defined as instantaneous rate of occurrence of the event at time t given that no event has occurred before time t [171]. Mathematically, the hazard function is defined as:

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T \leq t + \Delta t \mid T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} \tag{4.39}$$

which can also be written as

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\left[lnS(t)\right]. \tag{4.40}$$

So, the survival function (4.37) can be written as

$$S(t) = \exp(-H(t)) \tag{4.41}$$

where $H(t) = \int_0^t h(u)\,du$ denotes the cumulative hazard function.

The 'average' survival time is usually estimated by the median of the distribution. This is preferable to the expected value because of the skewness of the distribution and is given by solving the equation $F(t) = 1/2$.

The analysis of survival time data is largely studied. Statistical methods for estimation of the survival / hazard functions can be of three different types: non-parameteric, semi-parametric and parametric.

In this section, only parametric models and a continuous scale for survival time data will be considered; that is, models that require a specification of a probability distribution for the survival times. In particular, this means that one of the best known forms of survival analysis, the Cox proportional hazards model, is not considered because it is a semi-parametric model in which dependence on the explanatory variables is modelled explicitly but no specific probability distribution is assumed for the survival [165]. An advantage of parametric models, compared to the Cox proportional hazards model, is that it is easy to interpret, more efficient, accurate and there is a wider range of models with which to describe the data, including accelerated failure time models [171, 165].

Parametric Models Parametric models assume that the survival times or their logarithm of all instances in the data follow a particular theoretical distribution [171]. They can be rewritten specifying a direct relation between the survival time or its logarithm and the explanatory variables, just as a multiple linear regression model does. They are simple, efficient and effective in predicting the time to the event.

The commonly used distributions in parametric censored regression models are: normal, exponential, Weibull, logistic, log-logistic and log-normal. If the logarithm of the survival times of all instances follow these distributions, the problem can be analysed using the Accelerated Failure Time Models (AFT), these models assume that the variable can affect the time to the event of interest of an instance by some constant factor. It should be noted that if no suitable theoretical distribution is known, non-parametric methods are more efficient. [171]

The model parameters can be estimated by maximum likelihood. The likelihood function contains two components, one involving the uncensored survival times and the other making as much use as possible of information about the survival times which are censored [165].

Estimation Let N be the total number of instances such that c are censored and $N - c$ are uncensored observations, and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_P)^T$ be the set of model parameters.

For the j^{th} subject, the observed data are: t_j the survival time; δ_j the event indicator; and x_j a vector of explanatory variables. Let $t_1, \ldots, t_c$ denote the censored observations and $t_{c+1}, \ldots, t_N$ denote the uncensored ones. The contribution of the censored variables to the likelihood function is

$$\prod_{j=c+1}^{N} f(t_j, \beta). \tag{4.42}$$

For a censored observation, it is known that the survival time T is at least $t_j \quad (1 \leq j \leq c)$ and the probability of this is $Pr(T \geq t_j) = S(t_j)$, so the contribution of the censored observations to the likelihood function is

$$\prod_{j=1}^{c} S\left(t_j, \boldsymbol{\beta}\right). \tag{4.43}$$

The full likelihood is

$$\prod_{j=1}^{N} f\left(t_j, \boldsymbol{\beta}\right)^{\delta_j} S\left(t_j, \boldsymbol{\beta}\right)^{1-\delta_j}. \tag{4.44}$$

so, the log-likelihood function is

$$\begin{aligned} & \sum_{j=1}^{N} \left[\delta_j \log f\left(t_j, \beta\right) + \left(1-\delta_j\right) S\left(t_j, \beta\right)\right] \\ = & \sum_{j=1}^{N} \left[\delta_j \log h\left(t_j, \beta\right) + S\left(t_j, \beta\right)\right]. \end{aligned} \tag{4.45}$$

Note that these functions depend on the parameters of the probability distributions and on the parameters from the linear component $\mathbf{X}\boldsymbol{\beta}$. For several of the more commonly used probability distributions the requirements for generalized linear models are not fully met. Nevertheless, estimation based on the Newton-Raphson method for maximizing the likelihood provide good estimates [165].

The death density function $f(t)$, its corresponding survival function $S(t)$ and hazard function $h(t)$ for the most commonly use distributions are presented in Table 4.3.

Table 4.3: Density, Survival and Hazard functions for the distributions commonly used in the parametric methods in survival analysis

Distribution	PDF $f(t)$	Survival $S(t)$	Hazard $h(t)$
Exponential	$\lambda \exp(-\lambda t)$	$\exp(-\lambda t)$	λ
Weibull	$\lambda k t^{k-1} \exp(-\lambda t^k)$	$\exp(-\lambda t^k)$	$\lambda k t^{k-1}$
Normal	$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)$	$1 - \Phi\left(\frac{t-\mu}{\sigma}\right)$	$\frac{1}{\sqrt{2\pi}\sigma\left(1-\Phi\left(\frac{t-\mu}{\sigma}\right)\right)} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)$
Log-Normal	$\frac{1}{\sqrt{2\pi}\sigma t} \exp\left(-\frac{(\log(t)-\mu)^2}{2\sigma^2}\right)$	$1 - \Phi\left(\frac{\log(t)-\mu}{\sigma}\right)$	$\frac{1}{\sqrt{2\pi}\sigma t\left(1-\Phi\left(\frac{\log(t)-\mu}{\sigma}\right)\right)} \exp\left(-\frac{(\log(t)-\mu)^2}{2\sigma^2}\right)$

Exponential Distribution: Among the parametric models in survival analysis, the exponential model is characterized by only one parameter which is a constant hazard rate, λ. The occurrence of the event is assumed to be a random event independent of time. A larger value of λ indicates a higher risk and a shorter survival time period [171].

Weibull Distribution: The most common distribution in survival analysis is the Weibull distribution. It is characterized by two parameters, the scaling and the shape, $\lambda > 0$ and $k > 0$ respectively. The shape of the hazard function is determined using the shape parameter k, which provides more flexibility compared to the exponential model. If $k < 1$, the hazard function will be decreasing over time [171]. Similarly to the exponential model, a larger value of λ indicates a shorter survival time [171].

Normal and Log-normal Distribution: The normal distribution is characterize by two parameters, the mean μ and variance σ^2. [171].

Accelerated Failure Time Model The AFT models establishes a direct relationship between the predictors and the survival time, making its interpretation easier and are very similar to the conventional linear regression models [177]. In general, the AFT model can

be expressed as:

$$\ln(T) = \mathbf{X}\beta + \sigma\epsilon \tag{4.46}$$

where X is the covariate matrix, β represents the coefficients vector, $\sigma\,(\sigma > 0)$ is the unknown scale parameter, and ϵ is an error term which follows a similar distribution to $\ln(T)$.

Since the relationship between the logarithm of the survival time T and the covariates is linear in nature, it can be considered as a generalized linear model [171].

Typically, the parametric assumptions on ϵ can follow any of the distributions given in Table 4.3. In this case, the survival is dependent on both the covariates and the underlying distribution. Then, the only distinction of an AFT model compared to regular linear methods would be the inclusion of censored information in the survival analysis problem.

Model checking The model checking can be performed as in the case of the generalized linear models [165].

K - Nearest Neighbour

The K-nearest neighbour algorithm is one of the most popular classification algorithm methods and it can also be used for regression although this is not focused on this section. The algorithm performs a space partition in a pre-user-defined number of clusters (K) and by comparing a given test sample with a training sample. Each object is assigned to the class corresponding to the majority vote from its K nearest neighbours.

Suppose that the state space is $\left(\mathbb{R}^d, \|\cdot\|\right)$ where $\|\cdot\|$ is a reference distance. A common choice for $\|\cdot\|$ is the Euclidean distance but others can be used with similar properties.

Lets consider a fix $\mathbf{X} \in \mathbb{R}^d$. Given an i.i.d sample $S_n = \{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots, (\mathbf{X}_n, Y_n)\}$ in $\mathbb{R}^d \times \{0, 1\}$, where Y_i is the class label of $\mathbf{X}_i$. Given a new incoming observation $\mathbf{X}$, the goal is to predict its corresponding label Y.

The algorithm starts by re-ordering the S_n with respect to the distances $\| \mathbf{X}_i - \mathbf{X} \|$, such that

$$\| \mathbf{X}_{(1)} - \mathbf{X} \| \leq \cdots \leq \| \mathbf{X}_{(n)} - \mathbf{X} \|. \tag{4.47}$$

So, $X_{(m)}$ is the m-nearest neighbour of $\mathbf{X}$ and $Y_{(m)}$ is the corresponding label [181].

Given any integer k in $\mathbb{N}$, the principle of the nearest neighbour algorithm is to construct a decision rule based on the k-nearest neighbour of the $\mathbf{X}$: the S_n - measurable classifier $\Phi_{n,k}$ of the observation $\mathbf{X}$ is

$$\Phi_{n,k} = \begin{cases} 1 \text{ if } \dfrac{1}{k}\displaystyle\sum_{j=1}^{k} Y_{(j)} > \dfrac{1}{2} \\ 0 \text{ otherwise} \end{cases} \tag{4.48}$$

So, the decision is based upon a majority vote [181].

Longitudinal k-means algorithm

Consider a longitudinal study, collecting observations at multiple different times for each individual, thus defining individual trajectories. The longitudinal k-means algorithm is used to determine sets of homogeneous groups of trajectories.

Formally, let S be a set of the n subjects in the sample. For each subject i, denote by Y_{ik} the outcome of subject i measured at time k. The vector $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{it})$ is denoted trajectory.

K-means is an algorithm belonging to the Expectation Maximization (EM) class. It starts by randomly select k individuals from the data set and uses these as the initial means. Then it progresses by assigning each observation to the cluster where its distance to the center is minimal. The center of each cluster is then updated, from the mean of all observations within the cluster, and the process is iterated until some convergence criterion is attained. [182].

For measuring the 'closeness' the Euclidean distance is typically use which, in this context is defined as, for $i, j = 1, \ldots, n$,

$$D(Y_i, Y_j) = \sqrt{\frac{1}{t} \sum_{k=1}^{t} (Y_{ik} - Y_{jk})^2} \tag{4.49}$$

The Calinski and Harabasz criterion $C(g)$ can be used to determine the optimal number of clusters [182]. This criterion evaluates the clustering quality by combining the clusters between- and within variance. Let n_m be the number of trajectories in the cluster m; g the total number of clusters; $\bar{Y_m}$ the mean of the trajectory of the cluster m; $\bar{Y}$ the mean trajectory of the whole set S. The between-variance matrix is

$$\mathbf{B} = \sum_{m=1}^{g} n_m \left(\bar{Y_m} - \bar{Y}\right)^T \left(\bar{Y_m} - \bar{Y}\right). \tag{4.50}$$

The within-variance is

$$\mathbf{W} = \sum_{m=1}^{g} \sum_{k=1}^{n_m} \left(Y_{mk} - \bar{Y_m}\right)^T (Y_{mk} - Y_m) \tag{4.51}$$

For the optimal partition to have compact and well separated clusters it is desired a high value of between-variance and a low of within-variance. The Calinski and Harabasz criterion combines these two variability measurements in a ratio:

$$C(g) = \frac{Trace(B)}{Trace(W)} \times \frac{n-g}{g-l} \tag{4.52}$$

This is a heuristic measurement and so the proper way to use it is to compare clustering solutions obtained on the same data. This solutions may differ by the number of clusters or by the clustering method used.

Some clustering quality criteria commonly used are: some variation of the Calinski and Harabasz , Ray and Turi, Davies and Bouldin, BIC, BIC with correction for finite sample size, AIC and AIC with correction for finite sample size [182].

Naive Bayesian Classifier

A naive (or simple) Bayesian (NB) classifier is a probabilistic classifier which assumes that all features contribute equally and independently to the final decision [173]. It is a computational simple algorithm that can handle a data set with many attributes and thus widely-used in medical data mining. It combines a probability model with a decision rule.

The Naive Bayes conditional probability model assigns to each category of outcome variable Y ($Y \in \{C_1, \ldots, C_J\}$) and n independent features $\mathbf{X} = (X_1, \ldots, X_n)$ (X_i and X_j are independent for $i \neq j$ and $i, j \in \{1, \ldots, n\}$), it assigns to each instance the probabilities:

$$Pr(Y = C_k | X_1, \ldots, X_n) \tag{4.53}$$

for each of $k \in \{1, \ldots, J\}$ possible outcomes or classes C_k [183].

If n is large or if any of the X_i takes many values, then basing such a model on probability tables is infeasible. Using Bayes' theorem, the conditional probability can be decomposed as

$$Pr(Y = C_k \mid \mathbf{X}) = \frac{Pr(Y = C_k)\ Pr(\mathbf{X} \mid Y = C_k)}{Pr(\mathbf{X})} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}} = \text{posterior}. \quad (4.54)$$

Assuming that each feature X_i is conditionally independent of every other feature X_j for $j \neq i$, given the category C_k the joint model (4.54) can be written as:

$$\begin{aligned} Pr(Y = C_k \mid X_1, \ldots, X_n) &\propto Pr(Y = C_k)\ Pr(\mathbf{X} \mid Y = C_k) \\ &= Pr(Y = C_k, X_1, \ldots, X_n) \\ &= Pr(Y = C_k)\ Pr(X_1 \mid Y = C_k)\ Pr(X_2 \mid Y = C_k) \\ &\quad Pr(X_3 \mid Y = C_k)\ \cdots \\ &= Pr(Y = C_k) \prod_{i=1}^{n} Pr(X_i \mid Y = C_k)\,. \end{aligned} \quad (4.55)$$

One common decision rule is the maximum a posteriori or MAP decision rule which assign the observation to the class with the higher probability. The corresponding Bayes classifier is, therefore, the function that assigns a class label $\hat{y} = C_k$ for some k with the following the rule:

$$\hat{y} = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}}\ Pr(C_k) \prod_{i=1}^{n} Pr(x_i \mid C_k). \quad (4.56)$$

The class prior, $Pr\{Y = C_k\}$, can be estimated by assuming equiprobable classes, or by calculating an estimate for the class distribution in a training set.

The individual class-conditional marginal densities, $Pr\{X_i \mid Y = C_k\}$, can be estimated separately [18]. It can be estimated by assuming a distribution or by generating non-parametric models from the training set [184] .

If a component X_j is continuous, a common assumption is the Gaussian distribution, otherwise, an appropriate histogram estimate can be used. In the latter cases, multinomial and Bernoulli distributions are popular choices for the estimation proceeder.

Despite these rather optimistic assumptions, Naive Bayes classifiers often outperform far more sophisticated alternatives. Although the individual class density estimates may be biased, this bias might not hurt the posterior probabilities as much, especially near the decision regions [18].

In the context of survival analysis, the Naive classifier provide the probability of the event of interests as their outputs. The experimental results of using Bayesian methods on survival data show that Bayesian methods have good properties of both interpretability and uncertainty reasoning [171]. It has been used in several settings such as to make predictions in clinical medicine by estimating various probabilities from the data and integrating Bayesian methods with an AFT model by extrapolating the prior event probability to implement early stage prediction on survival data for the future time points [171].

It uses the Bayes theorem, which provides a link between the posterior probability and the prior probability, so that one can see the changes in probability values before and after accounting for a certain event [171].

One drawback of Naive Bayes method is that it makes the independence assumption between all the features, which can not be verified in some survival analysis problems [171].

Multilayer Perceptron Networks

Artificial Neural Networks were inspired by the biological neural systems. The 'simplest' artificial neural networks is composed by a collection of computational units denoted by 'neurons' that are connected based on weighted links, called edges, forming a network. Each neuron generates an output based on a certain kind of activation function and transmits throughout the link to another neuron.

Artificial Neural Networks can be used to define probabilistic models for regression and classification tasks by using network output to define the conditional distribution for one or more *targets*, $Y = C_k$, given the various possible values of a feature vector, $\mathbf{X}$ [185].

These algorithms are a popular alternative to conventional statistical models [186]. In particular feed - forward single hidden layer networks, with back - propagation training algorithms, are the most widely used and described here. They are effective in the analysis of complex data with non - linear trends and time - dependent covariates, and even high-order interactions [187].

A neural network is a two-stage regression or classification model, typically represented by a network diagram as the one in Figure 4.3 [18]. This diagram represents a network with p input (also named features or covariates) variables, M neurons in the hidden layer and K in the output layer.

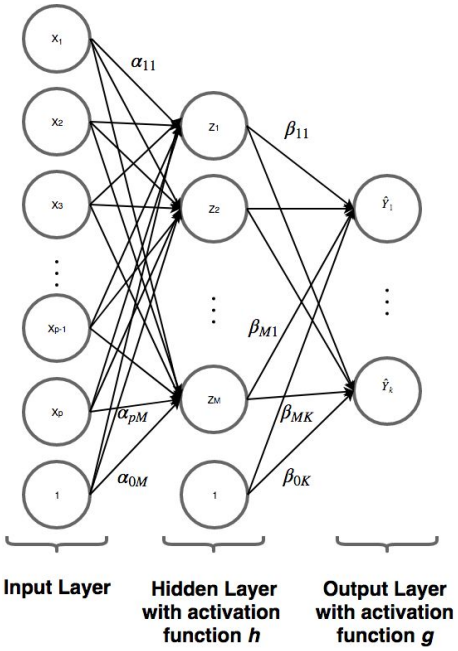


Figure 4.3: An example of the architecture of a feed - forward network having two layers of adaptive weighs.

For regression, usually $K = 1$ and there is only one output Y_1, but these networks can also handle multiple quantitative outputs. For a classification problem with K classes, there are K units, and the output of the k neuron of the last layer is the probability of the class k. There are K target measurements Y_k, $k = 1, \ldots, K$, each being coded as 0 or 1 for the k^{th} class [18].

The features Z_m are created from linear combinations of the inputs X_p, and then the target Y_k is modelled as a function of linear combinations of the Z_m:

$$
\begin{aligned}
Z_m &= h_m\left(\alpha_{0m} + \alpha_m^T \mathbf{X}\right), \quad m = 1, \ldots, M, \\
\hat{Y}_k &= g_k\left(\beta_{0k} + \beta_k^T \mathbf{Z}\right), k = 1, \ldots, K, \\
\hat{Y}_k &= f_k(\mathbf{X}) = g_k \circ h(\mathbf{X})
\end{aligned} \tag{4.57}
$$

where $Z = (Z_1, Z_2, \ldots, Z_M)$, $Z = (Z_1, \ldots, Z_K)$, $\alpha_m = (\alpha_{1m}, \ldots, \alpha_{pm})$, $\beta_k = (\beta_{1k}, \ldots \beta_{Mk})$ and the terms α_{0m} and β_{0k} are the bias capturing the intercept. The terms α_{ij} with $i = 0, \ldots, p$ and $j = 1, \ldots, M$ and β_{lu} with $l = 0, \ldots, M$ and $u = 1, \ldots, K$ are called weights.

The activation function $h(U)$ is usually chosen to be *sigmoid*: $h(u) = \dfrac{1}{1 + e^{-u}}$, although other functions may be used.

In regression problems, the output function $g(U)$ is usually the identity function and for classification problems with K classes the *softmax* function [18]:

$$
g(T_k) = \frac{\exp(T_k)}{\sum_{l=1}^{K} \exp(T_l)}. \tag{4.58}
$$

As any model, before finding the best set of parameters, it is necessary to define a measure of 'closeness' or 'fit' between the observed values and the estimate provided by the model.

So, considering the complete set of weights defined as

$$
\theta = \{\alpha_{im}, \beta_{jk}; i = 0, \ldots, p, m = 1, \ldots, M \text{ and } j = 0, \ldots, M, k = 1, \ldots, K\} \tag{4.59}
$$

The measurement of 'fit' depends on the task. If the task is regression, the commonly used measure is the sum-of-squared (error function):

$$
R(\theta) = \sum_{k=1}^{K} \sum_{i=1}^{N} (y_{ik} - \hat{y}_{ik})^2 \tag{4.60}
$$

where $y_{ik}; i = 1, \ldots, N; k = 1, \ldots, K$ is the target or observed data.

If the task is classification it is either used the squared error or cross-entropy (deviance):

$$
R(\theta) = -\sum_{k=1}^{K} \sum_{i=1}^{N} (y_{ik} \log(\hat{y}_{ik}))^2 \tag{4.61}
$$

where $y_{ik}; i = 1, \ldots, N; k = 1, \ldots, K$ is the target or observed data and the classification rule is $G(x) = argmax_k f_k(x)$, with $G(x)$ representing the class of observation x.

If the neuronal network used the softmax activation function and the cross-entropy error function, the model is exactly a linear logistic regression model in the hidden units, and all the parameters are estimated by maximum likelihood [18].

For networks having differentiable activation functions h and g, there exists a powerful and computationally efficient method, called error back-propagation for estimating the unknown weights. This approach minimize $R(\theta)$ by using gradient decent. Because of the compositional form of the model, the gradient can be easily derived using the chain rule for differentiation and by keeping track of quantities local to each unit in each a forward and backward sweep over the network [18].

Let $z_{mi} = h_m\left(\alpha_{0m} + \alpha_m^T \mathbf{X}_i\right)$ and $Z_i = (z_{1i}, z_{2i}, \ldots, z_{Mi})$, for the sum - of - squared (MSE) error function the algorithm is preformed as:

The learning rate γ_r is usually taken to be a constant, and can also be optimized by a line search that minimizes the error function at each update [18].

Algorithm 4.2 Back - Propagation algorithm

Require: Input: A dataset X with dimension $n \times p$ and a vector Y with dimension K.
Output: $\hat{\theta}$
Initialize network weights θ with random values (often near zero);
Set Epochs r= 0;
Initialize weights $\theta = \alpha\beta$ with random values;
Set a minimal value for $R(\theta)$: $R(\theta)_{mm}$;
Set temporary value for $R(\theta)$ that must be greater than $R(\theta)_{mm}$;
Set a maximum value for iterations - r_{max};
Set iteration counter as $r = 1$;
while $R(\theta) > R(\theta)_{mm}$ and $r < r_{max}$ **do**;
for each Observation $\mathbf{X}_i = [x_1, \ldots, x_p]$ and target $Y_i = [y_1, \ldots, y_k]$ **do** ;
Feed - Forward
for all Neurons in the network **do**
Calculate z_{mi} as in Equation (4.57);
Calculate $\hat{y}_{ik}$ as in Equation (4.57);
end for
Calculate $R(\theta)_i = \sum_{k=1}^{K} (y_{ik} - \hat{y}_{ik})$;
Back - Propagate:
for all Neurons in the network **do**
Calculate derivatives;
$\frac{\partial R_i}{\partial \beta_{km}} = -2 (y_{ik} - f_k(\mathbf{X}_i)) g_k' \left(\beta_k^T Z_i\right) Z_{mi}$;
$\frac{\partial R_i}{\partial \alpha_{ml}} = -2 \sum_{k=1}^{K} 2 (y_{ik} - f_k(\mathbf{X}_i)) \times g_k' \left(\beta_k^T Z_i\right) \beta_{km} h' \left(\alpha_m^T \mathbf{X}_i\right) X_{il}$;
Update Weights;
$\beta_{km}^{(r+1)} = \beta_{km}^{(r)} - \gamma_r \sum_{i=1}^{N} \frac{\partial R_i}{\partial \beta_{km}^{(r)}}$;
$\alpha_{ml}^{(r+1)} = \alpha_{ml}^{(r)} - \gamma_r \sum_{i=1}^{N} \frac{\partial R_i}{\partial \alpha_{ml}^{(r)}}$; Where γ_r is the learning rate;
end for
end for
Set $r = r + 1$
Calculate $R(\theta)_i = \sum_{k=1}^{K} R_i$
end whilereturn θ

The contribution of each variables in feed forward neural network can be assessed for determining the relative importance of independent variables to the final outcome. The connection weight method is one of the most used for this task. For each input node i, it calculates the sum of product of raw weights of the link from input node to hidden nodes ($W_{I->H}$) with the connection from hidden node to output nodes ($W_{H->O}$) for all input nodes:

$$R_I = \sum_{H=1}^{h} W_{I->H} W_{H->O} \tag{4.62}$$

The larger the sum for a given input node, the more the importance of the corresponding input variable [188].

Artificial neural network has been widely used in survival analysis with mainly three proposes:

- for predicting the survival time of a subject directly from the given inputs;
- for extending the Cox proportional hazard model;
- for obtaining a more flexible non-linear model by considering the censored information in the data using a generalization of both continuous and discrete time models [171].

Support Vector Machines Learning

SVM are the most popular and efficient classification and regression methods currently available. These algorithms apply simple linear methods in a high-dimensional feature space that is non-linearly related to the input space. Usually, all attributes are employed and non-overlapping partitions are generated.

Consider that our data is composed by N pairs $(\mathbf{X}_1, Y_1), ((\mathbf{X}_2, Y_2), \ldots, (\mathbf{X}_N, Y_N)$, with $\mathbf{X}_i \in \mathbb{R}^p$ and $Y_i \in \{-1, 1\}$ each indicating the class to which the point $\mathbf{X}_i$ belongs. The goal is to find the "maximum-margin hyperplane" that divides the group of points $\mathbf{X}_i$ for which $Y_i = 1$ from the group of points for which $Y_i = -1$, which is defined so that the distance between the hyperplane and the nearest point $\mathbf{X}_i$ from either group is maximized.

Any hyperplane can be written as the set of points $\mathbf{X}$ satisfying:

$$L = \left\{\mathbf{X} : f(\mathbf{X}) = \mathbf{X}^T\beta + \beta_0 = 0\right\} \tag{4.63}$$

where β is a unitary vector: $\parallel \beta \parallel = 1$ [18]. This hyperplane has the following properties:

- For any two points x_1 and x_2 lying in L, $\boldsymbol{\beta}^T(\mathbf{X}_1 - \mathbf{X}_2) = 0$, and hence $\beta^* = \frac{\beta}{\|\beta\|}$ is the vector normal to the surface of L.
- For any point $\mathbf{X}_0$ in L, $\boldsymbol{\beta}^T\mathbf{X}_0 = -\boldsymbol{\beta}_0$.
- The signed distance of any point $\mathbf{X}$ to L is given by

$$\begin{aligned} \beta^{*\mathbf{T}}(\mathbf{X} - \mathbf{X_0}) &= \frac{1}{\parallel \beta \parallel}\left(\boldsymbol{\beta}^T\mathbf{X} + \boldsymbol{\beta}_0\right) \\ &= \frac{1}{\parallel f'(\mathbf{X}) \parallel} f(\mathbf{X}). \end{aligned} \tag{4.64}$$

A classification rule induced by $f(\mathbf{X})$ is

$$G(\mathbf{X}) = sign\left[\mathbf{X}^T\boldsymbol{\beta} + \boldsymbol{\beta}_0\right]. \tag{4.65}$$

If the data is linearly separable, two parallel hyperplanes that separate the two classes of data can be selected, so that the distance between them is as large as possible. The region bounded by these two hyperplanes is called the "margin", and the maximum-margin hyperplane is the hyperplane that lies halfway between them. With a normalized or standardized dataset, these hyperplanes can be described by the following equations:

$$\mathbf{X}^T\boldsymbol{\beta} + \boldsymbol{\beta}_0 = 1 \tag{4.66}$$

and

$$\mathbf{X}^T\boldsymbol{\beta} + \boldsymbol{\beta}_0 = -1. \tag{4.67}$$

Geometrically, the distance between the two hyperplane is $\frac{2}{\|\boldsymbol{\beta}\|}$, so for the distance to be maximum, $\parallel \boldsymbol{\beta} \parallel$ must be minimum (Figure 4.4a). This optimization problem can be rephrased as:

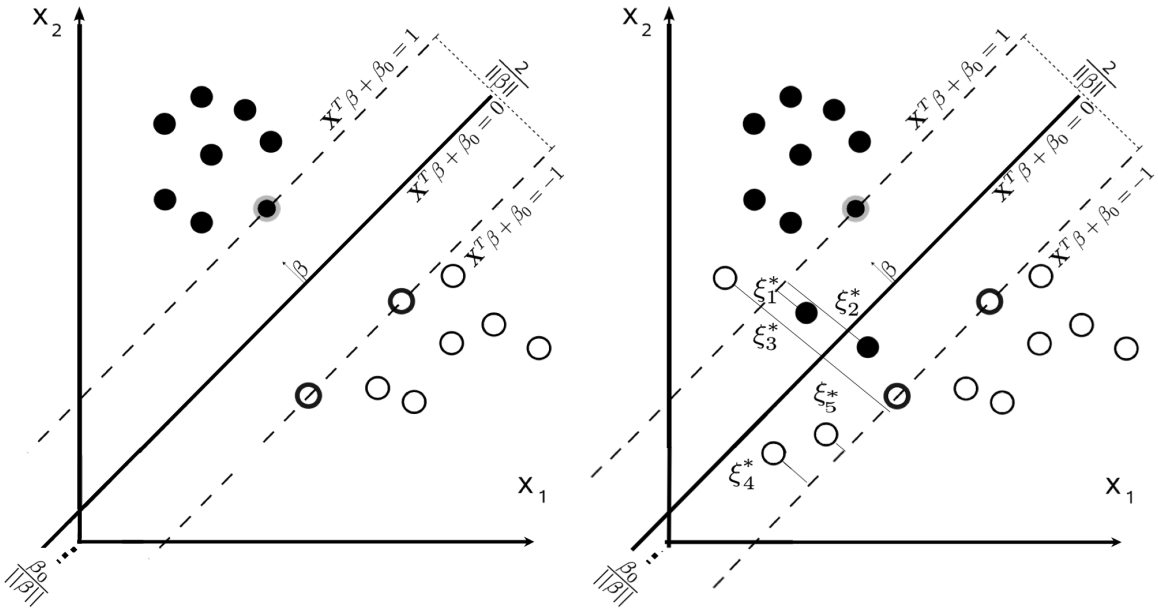
Problem 2

$$\min_{\boldsymbol{\beta}\boldsymbol{\beta}_0} \parallel \beta \parallel \ \textit{subject to}\ y_i\left(\mathbf{X}_i^T\boldsymbol{\beta} + \boldsymbol{\beta}_0\right) \geq 1, i = 1, \ldots, N. \tag{4.68}$$

When the classes have some overlapping, linear separating hyperplanes can still be a good solution. So, the problem can still be solved by maximizing the margin $\frac{2}{\|\boldsymbol{\beta}\|}$ allowing for some controlled cases to be on wrong side of the margin penalizing them.

Let $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_N)$ be the slack variables that define the amount by which data points are on the wrong side of the margin:

- if the point ξ_i^* lie on wrong side of the margin and is misclassified then $\xi_i > 1$
- if the point ξ_i^* is correctly classified but inside of the margin then $0 \leq \xi_i \leq 1$ (Figure 4.4b).



(a) Linearly separable classes with the solid line as the decision boundary and the dashed lines bound the maximal margin.

(b) Non-linearly separable classes with the solid line as the decision boundary and the dashed lines bound the maximal margin. The points label ξ_j^* are on the wrong side of their margin.

Figure 4.4: Maximum-margin hyperplane and margins for an SVM trained with samples from two classes. Samples on the margin are called the support vectors. [18]

If the point $\mathbf{X}_i$ rightly classified and in the right decision space then $\xi_i = 0$.

The optimization problem is a slight modification of the precious problem by introducing a introducing a penalty term:

Problem 3

$$\min \parallel \beta \parallel \textit{ subject to } \begin{cases} y_i \left(\mathbf{X}_i^T \boldsymbol{\beta} + \boldsymbol{\beta}_0 \right) \geq 1 - \xi_i \forall i, \\ \xi_i \geq 0, \sum_{i=1}^{N} \xi_i \leq constant. \end{cases} \tag{4.69}$$

Support Vector Machines are widely used mainly for classification problems even with survival data. It can be simply used considering only with recorded of patient experiencing the event in support vector regression or using the censored data with the constraint classification approach which imposes constraints in the SVM formulation for two comparable instances. Some authors have proposed a support vector regression for censored data, that can consider both uncensored and censored data, and takes advantage of the standard support vector regression and also adapts it for censored cases by using an updated asymmetric loss function [171].

In supervised methods one of the most common strategy for model validation is re-using efficiently the sample by preforming cross - validation.

Cross-validation consists of randomly dividing the data set in three independent parts: train, validation and test. The first dataset is use to fit the model, the validation set is used to estimate prediction error for model selection; the test set is used for assessment of the generalization error of the final chosen model [18]. Typical the training set will count 50% of the data, and the rest will be spit into equal parts for validation and testing.

4.1.2 Model Validation and Selection - Some General Considerations

The model validation and selection is an important component of model fitting because model misspecification may cause severe bias and even lead to incorrect inference [189]. Since models are only approximations and a perfect fit may be too much to ask for, the problem becomes assessing the model's performance on approximating the true model. This problem can also be formulated as how to select a model amongst different models for a given research question which, in practice, is more important that the actual method for estimating the parameters. The basic idea is that, even if the method is useless for the intent for which it was specified, the proper model selection will reveal this weak approach and show that other methods are preferable.

There are two main approaches for model validation: analytically through goodness-of-fit measures or by efficient sample re-use (cross-validation or bootstrap) [18].

Unfortunately, there are no formal standards for how to evaluate the quantitative goodness-of-fit of models to data, either visually or numerically resulting in a considerable number of methods. While there are some subtle and perhaps controversial issues involved in the evaluation of goodness-of-fit, there are many simple conventions that are quite uncontroversial; those will be described in this subsection [190].

The goodness-of-fit test compares the model fit with the data that represents the fit of the most complex model possible - the saturated model, having a parameter for each observation [164]. The most common measures are described in [168] and in [190] (some of the measures only work for specific models):

- Standard χ^2 goodness of fit: $\chi^2 = \sum \frac{(O-E)^2}{E}$ - the difference between model predictions and data on each condition mean is squared and divided by the model prediction;
- Residual sum of squares: $\hat{\epsilon}^T \hat{\epsilon}$ - measures how well the model fits in an absolute sense;
- Coefficient of determination (or percentage of explained variance): R^2 - measures how well the models fits in a relative sense. It measures the proportion of variation in the outcome that is accounted for by the predictors. Measures the extent to which the model accounts for the observed data. Only available for the linear model. Some extensions have been proposed for, at least, logistic regression;
- Mean Squared Error or Root Mean Squared Error: $MSD = \frac{\sum_{i=1}^{k}(m_i - d_i)}{k}$ and $RMSD = \sqrt{MSD}$, where m_i is the model mean for each point i and d_i is the data mean for each point i and k is the number of points i being compared - this measure gives more emphasis is placed on points that do not fit well than on points that do fit well;
- AIC: $AIC = -2\left(\ln\left(\text{likelihood}\right) - k\right)$ where k is the number of parameters in the model - The optimal model is the one that minimizes this quantity [164].
- BIC: $BIC = \ln\left(n\right)k - 2\ln\left(\text{likelihood}\right)$ where n is the number of points and k the number of models parameters. The preferred model should be the one the minimizing this quantity.

For several reasons, goodness-of-fit indexes whose contribution is just fit versus does not fit, such as the χ^2 test, should be avoided for assessing the fit of theoretical models to data [190]. They provide positive evidence only by accepting the null hypothesis and favour data sets with low number of cases and with increased noise it is less likely to obtain a "does not fit" outcome. Moreover, models have heuristic and summative value even though a model does not capture all the non-noise variance in the data, it may capture many important aspects. If a model is currently the only model of a phenomenon or best model, then one should not discard it as completely false because one aspect of the data is not yet captured [190, 191].

On the contrary, if we have a very large sample, the statistical test will almost certainly be significant. The low p-values might be an artefact of large-sample sizes and, if this is the case, we will wrongly reject the model, even if it actually describes the data quite well - the Cried Wolf effect [191].

A saturated model that specifies all possible paths between all variables always fits the data perfectly, but it is just as complex as the observed data and a trade-off between the two must be achieved [191]. The goal is to use a goodness-of-fit index that does not depend on the sample size or the distribution of the data. In fact, simulations have shown that most goodness-of-fit indices still depend on sample size and distribution, but the dependency is much smaller than that of the routine χ^2 test [191].

Other perspective of models quality is its generalization performance which relates to its prediction capability on providing reliable estimates in independent test data [18]. So, typically, data is divided in training and testing whose behaviour according to the model complex is represented in Figure 4.6. The training error tends to decrease with the increase of model complexity. However, the model becomes too proficient in describing the training data and will not generalize well (i.e., have large test error) - this phenomenon is called over-fitting. In that case the models predictions will have a large variance. On the contrary, if the model is not complex enough, it will underfit and may have large bias, again resulting in a poor generalization [18].

It is well known that models with low bias in parameters estimation have a higher variance of the parameter estimates across samples, and vice versa. High bias can cause a model to miss the relevant relations between features and target outputs - underfitting phenomenon. The variance measures the model sensibility to small fluctuations in the data, so a high variance can cause overfitting describing the random noise in the training data. So, interplay between bias, variance and model complexity must be carefully analysed.

Let's consider a response variable Y and X a set of predictors such that $Y = f(X) + \epsilon$ where $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma_\epsilon^2$. The expected prediction error of a regression model fit $\hat{f}(X)$ at a point $X = x_0$ using squared-error loss, can be written as:

$$Err(x_0) = E[(Y - \hat{f}(x_0))^2 | X = x_0] \tag{4.70}$$

For simplicity, the conditional term will be dropped and so the squared-error loss can be written as:

$$\begin{aligned} Err(x_0) &= E[Y^2] - 2E[\hat{f}(x_0)]E(Y) + E[\hat{f}^2(x_0)] \\ &= Var(Y) + f^2(x_0) - 2E[\hat{f}(x_0)]f(x_0) + Var(\hat{f}(x_0)) + E(\hat{f}(x_0))^2 \\ &= Var(Y) + [E(\hat{f}(x_0)) - f(x_0)]^2 + Var(\hat{f}(x_0)) \\ &= \sigma_\epsilon^2 + Bias^2(\hat{f}(x_0)) + Var(\hat{f}(x_0)) \\ &= \text{Irreducible Error} + \text{Bias}^2 + \text{Variance}. \end{aligned} \tag{4.71}$$

The 'Irreducible Error' is the variance of the response variable around its true mean $f(x_0)$, and cannot be avoided no matter how well we estimate $f(x_0)$, unless $\sigma_\epsilon^2 = 0$. The second term is the squared bias, the amount by which the average of the estimate differs from the true mean; the last term is the variance, the expected squared deviation of $\hat{f}(x_0)$ around its mean. Typically, the more complex the model $\hat{f}$, the lower the (squared) bias but the higher the variance [18].

Figure 4.5 shows the bias–variance trade-off schematically. Considering a linear model as an example, the model space is the set of all linear predictions from p inputs and the black dot labelled "closest fit" is $\mathbf{X}\beta$. The blue-shaded region indicates the error σ_ϵ with

which the truth in the training sample is represented. Also shown is the variance of the least squares fit, indicated by the large yellow circle centred at the black dot labelled 'closest fit in population'. If a model is fit with fewer predictors, or the coefficients are shrinked towards zero, it would result in a "shrunken fit". This fit has an additional estimation bias, due to the fact that it is not the closest fit in the model space but smaller variance. If the decrease in variance exceeds the increase in (squared) bias, then this is worthwhile [18].

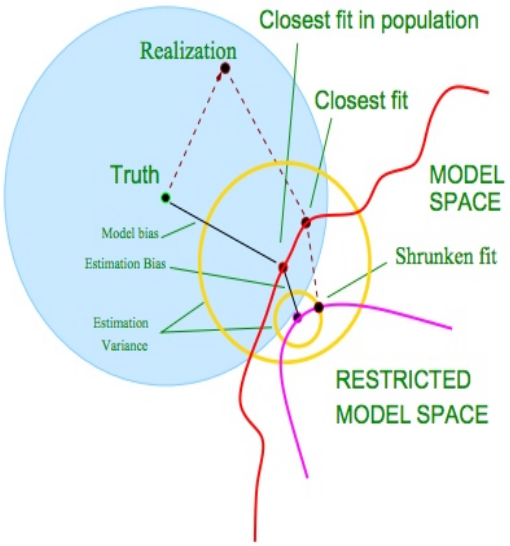


Figure 4.5: Schematic picture of the behaviour of bias and variance [18].

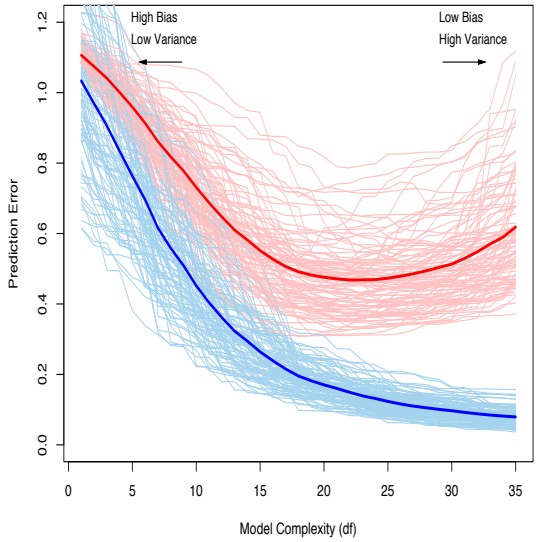


Figure 4.6: Behaviour of test sample and training sample error as the model complexity varies [18].

The selection of the explanatory variables in regression models, is a particular case of model selection but with the same considerations about the bias-variance trade-off.

It must take into consideration that the more complex the model, the more capable it is to capture the underlying variance but the less able it is to predict the response to new data. That is, adding variables turns the model specialist in the particular observed data set used to estimate the parameters. Indeed, by using as many parameters as observations the fit can be made perfectly [17].

Traditionally, stepwise methods for variables selection focus on iteratively adding or eliminating variables. These method have been rising some controversy; two of the most common argument is highlighted:

> People sometimes think that if a coefficient estimate is not significant, then it should be excluded from the model. We disagree. It is fine to have non significant coefficients in a model, as long as they make sense [169].

and

> Unfocused search throughout many possible models increase the likelihood of capitalizing on change and thereby finding a model that represents only a spurious relationship. It is our experience and strong belief that better models and a better understanding of one's data result from focussed data analysis, guided by substantive theory [192].

Models identified by stepwise methods have an inflated risk of capitalizing on chance features of the data. They often fail when applied to new datasets but also, are rarely tested in this way [192].

Frank Harrell, in [193] compiled a list of problems with automated stepwise model selection procedures:

1. It yields R^2 values that are badly biased to be high.
2. The F and chi-squared tests quoted next to each variable on the printout do not have the claimed distribution.
3. The method yields confidence intervals for effects and predicted values that are falsely narrow [194];
4. It yields p-values that do not have the proper meaning, and the proper correction for them is a difficult problem.
5. It gives biased regression coefficients that need shrinkage (the coefficients for remaining variables are too large [195];).
6. It has severe problems in the presence of collinearity.
7. It is based on methods (e.g., F tests for nested models) that were intended to be used to test prespecified hypotheses.
8. Increasing the sample size does not help very much [196];.
9. It allows us not to think about the problem.
10. It uses a lot of paper.

As an alternative, variables selection should be regarded as model selection and should use the same methods always grounded in solid theory.

4.2 Modelling the Reporting - Delay Distribution and Incidence

The statistical problems involving reporting delay with right-truncated data has a history stretching back at least 50 years being Sartwell's, with the study of incubation periods using log-normal distributions, one of the earliest reference. Most recently, Bayesian methods have been used to handle situations in which infection times are unobserved and data mining techniques were also addressed in contexts of survival time of cancer patients, and reporting delay of HIV-AIDS. Similar problems involving right-truncation data arises from analysis of AIDS latency where the time of onset of AIDS is truncated since only those HIV infected individuals who develop AIDS within the truncation interval are observed. [8, 197, 186, 24]. It is precisely the HIV-AIDS context that has been providing a great impulse to the statistical treatment of right-truncated with focus in obtaining valid estimates of recent incidence, when case ascertainment or reporting is subject to delays [198].

In this section it will be reviewed the key publications on statistical inferences consisting of estimating the lag distribution between an identification of an event and its reporting, and the size of the random sample of the diagnosed cases using the observed (reported) cases [8]. The full literature review summary table of all key publications in these settings as well as related work is given in Appendix F.

The most common approaches arise from joint estimation of the reporting delay and incidence, which can be grouped into two major approaches whose focus is one of the two distributions. The subsection 4.2.1 overviews key publications of the two most common joint modelling approaches in this context focusing of chosen approaches, objectives, estimation techniques and strengths and limitations; subsection 4.2.2 overviews key publications of separated modelling approaches and subsection 4.2.3 gives a brief review of the approach for the underlying statistical problem of the two most common approaches of joint modelling: to make inferences about a stochastic process (estimating the reporting delay distribution

as well as the random sample size) for which realizations are right-truncated in an interval defined by the analysis; if a diagnosis takes place at time x and the analysis at time x^*, only cases that have reporting delays less than time $x^* - x$ will be considered.

4.2.1 Joint Modelling

A general formulation involving joint estimation of the incidence distribution as well as the reporting delay distribution, based on two sub-models linked through specific association parameters, is the starting point with Harris [198]. Under this approach, several methods have been developed for analysing right truncated data which can be largely grouped into two major categories [163]: one is to treat the reporting delay as the time between the occurrence of the event and its report and use survival analysis methods to estimate the delay distribution, and the other considers the observed cases as a discrete outcome and uses models for discrete responses such as the Poisson and Multinomial to model the delay distribution. The former typically focuses on inference about the reporting delay distribution, while the latter concentrates on the disease incidence [163].

Under the first approach, consider fitting a regression model based on a discrete reverse - time proportional hazard function. Brookmeyer and Liao in [199] proposed a convenient method for the analysis of grouped data that does not involve the incidence function [198]. This popular model is simple when implemented with a complementary log - log link, inducing a relationship between the Cumulative Distribution Function (CDisF) of the delay distribution of the form:

$$F\left(t \mid \mathbf{X}\right) = F_0\left(t\right)^{\exp\left(\beta^T \mathbf{X}\right)}$$

where $F\left(t \mid \mathbf{X}\right)$ is the CDisF of delays in a sub-population with characteristics $\mathbf{X}$ and $F_0\left(t\right) = F\left(t \mid 0\right)$ [198]. As Brookmeyer and Liao had noted, this model is not a proportional hazards model because when $\beta^T \mathbf{X} > 0$, the ratio of hazard functions increases monotonically from 0 to 1, thus converging at long delays [198]. This model, can be implemented as a generalized linear model, and has been largely used to study the effect of several factors on the delays in AIDS such as area of residence, exposure category, seasonal variation [119]. It relies on a result for right-censored data under non-informative censoring whose validity for right-truncated data requires further investigation [8]. Kalbfleisch, in [200], also fitted a regression model based on a reverse - time hazard function but applied to AIDS latency distribution. This approach expressed the reverse time hazard for both discrete and continuous times. For discrete times, he developed a full likelihood while for continuous times, a partial likelihood that can only be applied to proportional hazards model and accommodates only right truncated data. Nonetheless, the models may be used to assess the quasi - stationary of lag distributions and to provide flexible representations of covariate effects for truncated data. A similar approach was used by Lagakos to obtain non-parametric estimates of AIDS latency distributions [201].

The reverse - time proportional hazard function approaches generally imply a transformation of the data and the interpretation of the results is not simple. We notice that the approach was mainly tested with the Centers for Disease Control and Prevention (CDC) data who collects data from the U.S. epidemic.

Under the second approach, that of discrete outcome models, Harris described the joint probability function for reporting delay and incidence under the assumption that the cross - classified counts are independent and Poisson distributed. He also assumes that the likelihood function can be written as a product of the conditional and marginal likelihoods [121]. He described categorical and mixed categorical / continuous - time models for reporting delay while for incidence described only categorical and continuous - time. The incidence of HIV cases was obtained by back-calculation over the incidence of AIDS cases.

This approach can accommodate both truncated and censored data and the obtained back-calculated results were consistent with the used serological data. However, the method imposes constraints on the distributions that may be satisfied only by some type of regression models. Moreover, the counts of cases are very unlikely to be Poisson and if cases are reported in batches the independent multinomial assumption does not capture the pattern.

Pagano, in 1994, discussed an extension of a discrete outcome regression model that can be applied not only to truncated data but also to censored. This approach treats truncation, as well as censoring, as a type of missing data and discussed the application of missing-data techniques to simplify estimation. They augmented the incomplete data using the most direct missing - data technique in this setting: the EM algorithm. The algorithm iterates between the E-step, which augments the observed portion of the sample by an estimate of the unobserved portion due to truncation, and the M-step, which maximizes the log-likelihood of the augmented data, until a convergence criterion is reached [8].

Green, in 1998, described adjustments to be made to the number of AIDS cases considering delays in reporting, the lack of HIV-exposure information for some cases, and future diagnoses of AIDS-defining opportunistic illnesses among persons reported with AIDS under severe immune - suppression. The cases were separated into clusters that are small enough for within-cluster reporting delays to be fairly homogeneous but large enough to yield precise estimates of their reporting delay distributions. First, he evaluated a set of covariates to determine which of their levels are associated with reporting delay. Second, the reporting delay distribution is estimated separately for each group of cases by the cross-classification of levels identified during the first stage, and groups with similar estimated delay distributions are combined. The covariates are selected using binary response regression models. This approach takes the context of the reporting institution into consideration and allow variations in AIDS incidence trends by mode of exposure to HIV, sex, race/ethnicity and geographic region which are in qualitative agreement with reported variations in HIV prevalence rates and changes in HIV prevalence over time. This analysis use a very short period of time and assumed that the reporting delay does not vary over six-year period within groups for which separate estimates are made. Moreover, a large uncertainty in the exposure distributions is described [120].

Amaral et al., in 2000 and 2005, applied a back-calculation method for obtaining forecasts of the Portuguese incidence data accounting for reporting delay and under-reporting. The method used to estimate the reporting delay distribution is based on Poisson regression and involves cross-classifying each reported case by calendar time of diagnosis and reporting delay, similarly to the Brookmeyer approach in [202]. The adjusted AIDS incidence data are then used to obtain short-term projections and lower bounds on the size of the AIDS epidemic using a Weibull incubation period distribution. The EM algorithm is used to obtain maximum-likelihood estimates when the density of the infection times is parametrized as a step function. Recent discussions suggest that the back-calculation method is gradually becoming less appropriate for reliable incidence and prevalence estimates, as it does not take into account the effect of treatment. Only the lower bound of the epidemic was defined. The authors pointed as limitations, the unknown incubation period distribution and the inaccuracy in the observed disease incidence over time, the assumptions made relative to the reporting delay distribution were very unlikely and assumptions on under-reporting percentage and treatment effect were not taken into consideration [203, 204].

Midthune et al., in 2005, generalized the jointly Harris model to include, along with reporting delay, data correction to the cancer registries data . The reporting delay component of the model which, as its predecessors, is specified for grouped data, parametrizes the probability of a given delay in terms of the discrete hazard, using a complementary log–log link function, though the model is not a generalized linear model [198]. They divided the registries population into sub-populations based on the usual subgroups used for report-

ing, that is, one for each combination of levels of the following variables: year of diagnosis (17 years), gender, race (for melanoma, they used only whites, because melanoma is rare among other racial groups), and 5 - year age groups. All reporting models assumed that the reporting process was relatively stable. Although the reporting-delay model with a random reporting-year effects fitted the data much better than the non-random model, it still exhibited some lack of fit. For simplicity, it was assumed that the random effects were normally distributed and independent; alternatively, one could allow the random effects to be correlated or use non-parametric methods to estimate the distribution of the effects. This model allows correlation within reporting year and it would be useful to also allow correlation within diagnosis year [205]. Although parametric assumptions allow for the estimation of the distribution of the reporting delays, the results are extremely imprecise and depend strongly on the assumptions [117]. Moreover, any effect, no matter how tiny, can produce a small p-value if the sample size or measurement precision is high enough [206].

As an alternative, non-parametric data mining techniques have been used in cases where similar patient records and related symptoms were used [172].

4.2.2 Separate / Partial Modelling

Ideally, the reporting delay can be handled by jointly modelling and estimating the incidence and reporting delay distributions. Unfortunately, this joint estimation process is not trivial to perform due to missing information in the notification forms [116]. To overcome those difficulties, separate approaches have also been proposed, one of the first authored by Barnard in 1999.

Barnard, in [116], reviewed three applications of the well known Rubin's imputation method that are directly relevant for medical studies. One of them consists of the estimation of the survival time after AIDS diagnosis in the presence of reporting delay, using data from the CDC AIDS surveillance system. The emphasis of this review was on the building of imputation models (i.e. the first step), which is the most fundamental aspect of the method. This approach allows dealing separately with the reporting delay and with the survival analysis, thus avoiding the difficulties of joint estimation. One can first concentrate on the modelling of the survival time without worrying about the reporting delay, and then estimate the last distribution using the available information. Then, it is suggested to multiply impute the delayed cases and proceed with the second and third steps of Rubin's method to be able to infer about the survival time. However, if the imputation model is seriously flawed in terms of capturing the missing-data mechanism, then so will be any analysis based on such imputations. It is less efficient than joint modelling of the missing-data mechanism and the substantive analysis, in terms of both statistical efficiency (e.g. avoiding the reliance on a finite number of imputations) and computational efficiency (e.g. avoiding simulation via the use of the EM algorithm).

Xia, in 2015 in [163], developed a distribution-free approach to allow separate modelling of the incidence and reporting delay but with joint inference for both modelling components, based on functional response models. This separate approach allows the accommodation of a sub samples of data to model the reporting delay. The model is fitted using a set of weighted generalized estimating equations adapted to the functional response models ir order to provide consistent parameter estimates and valid inferences. The authors discussed inferences about projections of future disease incidence to help identify significant shifts in temporal trends. Simulated data provided good performances, even for relatively small sample sizes .

In the same year, Noufaily in [198] focus just on the reporting delay and developed a model for describing the hazard to gain a better understanding of the reporting process being the recent incidence a possible covariate. They also investigated the effects of tem-

poral factors such as calendar time, season and recent incidence on the reporting delay. The authors use a continuous time spline - based model for the hazard of the delay distribution along with an associated proportional hazard model. This approach allows a natural investigation of covariates effects, with a relative simple hazard interpretation and also the inclusion of long delays which are often excluded from the analysis. The major disadvantage of the approach is that it is more cumbersome than the generalized linear model method. Noufaily also noticed that the estimates of the hazard at long delays are sensitive to irregularities in the data.

Table 4.4 presents a brief comparison of the main approaches, described above, for the estimation of the reporting delay and incidence.

Table 4.4: Comparison of reporting delay estimation models

	Joint Modelling		Separate Modelling
	Reverse time hazard	Discrete outcomes	
Focus	Reporting delay	Incidence	Reporting delay and / or incidence
Data needed	Case base data	Aggregated data	Case based data / Aggregated data
Strengths	Allows the visualization the hazard in continuous time and study the effect of covariates in a natural fashion. Can describe and characterize the entire delay distribution, accommodating the long delays.	This approach accommodates truncation without transforming the data. Can be implemented with standard statistical packages.	Provides more flexibility for modelling the sub-models, possibly with different samples, allowing more complex approaches.
Weaknesses	Generally implies data transformation and interpretation of the results is not simple.	Need parametric assumptions, which may yield biased inferences if data fail to meet the posited distributional assumptions [163].	May need parametric assumptions on the incidence component.
Key publications	[201], [202], [122], [207], [198]	[121], [208], [8], [116], [209], [120], [203], [204], [210], [205]	[197], [163]

4.2.3 Data - structure and notations

Let t_0 represent the earliest time date and t_n the latest time date for which an HIV-AIDS case(s) is known or the end of the observation period is available. Consider that the time interval $[t_0, t_n]$ can be divided into n units of equal length of the form $[t_{j-1}, t_j]$ which are indexed by non-negative integers $j = 1, \ldots, n$. For simplicity, and without loss of generality, we will consider the time interval $[t_{j-1}, t_j]$ represented only by t_j. Let d be the reporting delay. That is, if an HIV - AIDS case is diagnosed at the time period t_j it only gets reported until d time periods latter; the reporting time is thus $t_j + d$ for $j = 1, \ldots, n$. This case will only be observed if $t_j + d \leq t_n$, or equivalently $0 \leq d \leq t_n - t_j$.

Let $\mathbf{D}$ be a random variable of reporting delays taking values $0 < d_1 < d_2 < \ldots < d_i < \ldots$. The distribution of $\mathbf{D_j}$ conditional to a covariate vector $\mathbf{Z_j}$ has CDistF:

$$F_j\left(d_i|\mathbf{Z_j}\right) = Pr\left\{D \leq d_i|\mathbf{Z_j}\right\} \tag{4.72}$$

We assume that d_m is the longest reporting delay that can be reliably estimated and so $F_j\left(d_m|\mathbf{Z_j}\right) = 1$. Under this assumption, cases diagnosed prior to $t_n - d_m$ are all reported by time t_n, while only a proportion of those whose diagnose occurs between $t_n - d_m$ and t_n ($t_n - d_m < t_j < t_n$), is observed (with reporting delay $d < t_n - t_j$) [163].

All HIV-AIDS cases can be cross - classified by their date of diagnosis and the length of their reporting delay: let Y_{ij} be the number of HIV-AIDS cases diagnosed at t_j with reporting delay d_i. The incidence in time t_j is denoted by Y_{+j} and given by:

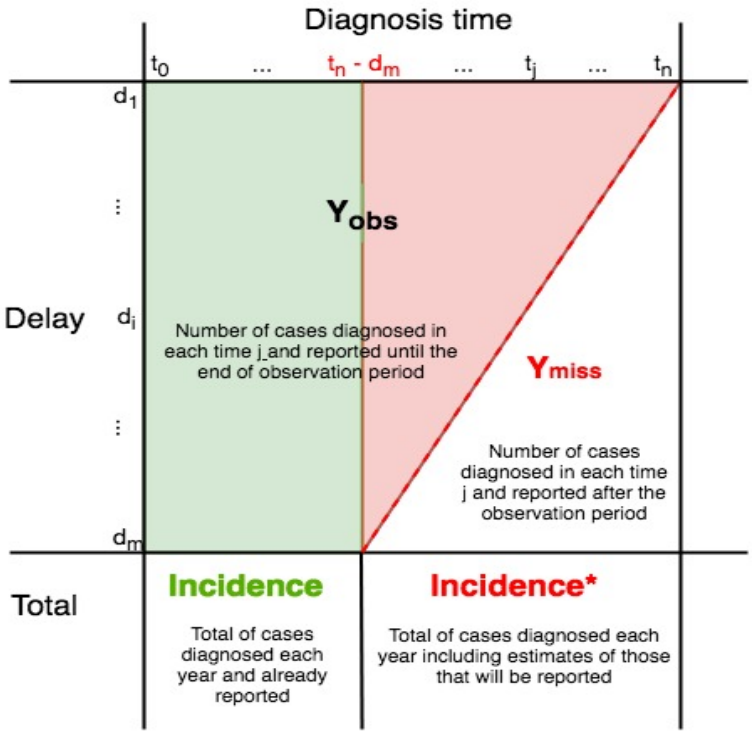


Figure 4.7: Truncation mechanism in reporting cases based on [211]

$$Y_{+j} = \begin{cases} Y_{+j}^{C} = \sum_{i=0}^{m} Y_{ij} = & \text{if } t_0 \leq t_j \leq t_n - d_m \\ Y_{+j}^{Tr} = \sum_{i=0}^{t_n - t_j} Y_{ij} & \text{if } t_n - d_m + 1 \leq t_j \leq t_n \end{cases} \tag{4.73}$$

Shown in Figure 4.7 is the typical representation of the data structure and the formulation of this problem: making inference about an unobservable random variable with right-truncated values defined by the observation period.

Joint modelling of the Reporting Delays as a Survival Time

Since D is, by definition, the time that a diagnosed HIV-AIDS case gets reported to the surveillance system, it can be viewed as a 'survival time' and survival methodology can be employed for the estimation of its distribution.

Let D be a discrete random variable taking values $0 \leq d_1 < d_2 < \ldots < d_i$ with probabilities conditional to a covariate vector $\mathbf{Z_j}$

$$f_j\left(d_i \mid \mathbf{Z_j}\right) = f_{ij}\left(\mathbf{Z_j}\right) = Pr\left\{D = d_i \mid \mathbf{Z_j}\right\} \tag{4.74}$$

Let $F_j\left(d_i \mid \mathbf{Z_j}\right)$ be the cumulative probability function

$$F_j\left(d_i \mid \mathbf{Z_j}\right) = F_{ij}\left(\mathbf{Z_j}\right) = Pr\left\{D \leq d_i \mid \mathbf{Z_j}\right\}. \tag{4.75}$$

Moreover, define the probability of a case being reported d_i units of time after diagnosis time t_j, given that it is reported within d_i units of time, as

$$g_j\left(d_i \mid \mathbf{Z_j}\right) = g_{ij}\left(\mathbf{Z_j}\right) = \frac{f_{ij}\left(\mathbf{Z_j}\right)}{F_{ij}\left(\mathbf{Z_j}\right)} = Pr\left\{D = d_i \mid D \leq d_i, \mathbf{Z_j}\right\}. \tag{4.76}$$

The function g_j is the reverse time hazard function and as Kalbfleisch pointed in [122], only the reverse time hazards are strictly identifiable with truncated data.

Under the assumption of a maximum delay d_m, (4.75) can be written as:

$$F_{ij}\left(d_i \mid \mathbf{Z}_j\right) = \begin{cases} Pr\left\{D_j \leq d_i \mid \mathbf{Z_j}\right\} & \text{if} \quad t_n - d_m + 1 \leq t_j \leq t_n \\ 1 & \text{if} \quad t_0 \leq t_j \leq t_n - d_m \end{cases}, 0 \leq d_i \leq d_{m-1}. \quad (4.77)$$

Note that if a case (diagnosed between $t_n - d_m + 1 \leq t_j \leq t_n$) has a reporting delay less than d_i (with $0 \leq d_i \leq d_{m-1}$) , it implies **not having** a reporting delay of $d_{(i+1)}$ given that it is reported until $d_{(i+1)}$ unit times later, nor $d_{(i+2)}$ given that it is reported until $d_{(i+2)}$ unit times later, ... Hence (4.77), and as Kalbfleisch in [122] demonstrated and considering (4.76), can be written as:

$$F_{ij}\left(d_i \mid \mathbf{Z}_j\right) = \begin{cases} \prod_{k=d_{(i+1)}}^{d_m} \left(1 - g_{kj}\right) & \text{if} \quad t_n - d_m + 1 \leq t_j \leq t_n \\ 1 & \text{if} \quad t_0 \leq t_j \leq t_n - d_m \end{cases}, 0 \leq d_i \leq d_{m-1}. \quad (4.78)$$

To model g_{ij}, let $Y_{+_k j}$ be the total number of cases diagnosed in t_j and reported until d_k and Y_{kj} the total number of cases diagnosed in t_j with reporting delay exactly equal to d_k ($1 \leq d_k \leq d_m$). Then

$$Y_{kj} \mid Y_{+_k j}, \mathbf{Z_j} \overset{iid}{\sim} Bi(Y_{+_k j}, g_{kj}\left(\mathbf{Z_j}\right)), \text{ with } \begin{cases} 0 < d_i < d_m & \text{if} \quad 0 < t_j < t_n - d_m \\ 0 < d_i < t_n - t_j & \text{if} \quad t_n - d_m + 1 < t_j < t_n \end{cases}$$

where Bi represents a binomial distribution. A natural class of regression models for (4.76) is [122]:

$$\psi\left(g_{ij}\right) = \boldsymbol{\beta_0} + \mathbf{X}_j^T \boldsymbol{\beta}_x \quad \text{with} \quad \begin{cases} 0 < d_i < d_m & \text{if} \quad 0 < t_j < t_n - d_m \\ 0 < d_i < t_n - t_j & \text{if} \quad t_n - d_m + 1 < t_j < t_n \end{cases} \quad (4.79)$$

where ψ is a function of $[0, 1]$ onto $[-\infty, \infty]$, $\boldsymbol{\beta}$ is a vector of the regression parameters and $\boldsymbol{\beta_0}$ is the Reverse Time Hazard Function (r.t.h.) baseline [122].

Two natural choices for link functions $\psi\left(\cdot\right)$ are the logit function:

$$\psi\left(u\right) = \log\left(\frac{u}{1-u}\right) \quad (4.80)$$

and the complementary log-log, a particular popular choice :

$$\psi\left(u\right) = \log\left\{-\log\left(1-u\right)\right\} \quad (4.81)$$

For the model in (4.81), $F_{ij}\left(d_i|\mathbf{X_j}, \boldsymbol{\beta}\right)$ can be written as:

$$F_{ij}\left(d_i \mid \mathbf{X_j}, \boldsymbol{\beta}\right) = F_0\left(d_i \mid \boldsymbol{\beta}\right)^{\exp\left(\mathbf{X}_j^T \boldsymbol{\beta}_x\right)} \text{ with } F_0\left(d_{ij} \mid \boldsymbol{\beta}\right) = \prod_{k=d_{(i+1)}}^{d_m} \exp\left(-\exp\left(\boldsymbol{\beta_0}\right)\right) \quad (4.82)$$

$$\text{where } 0 \leq d_i \leq d_{m-1}, \quad t_n - d_m \leq t_j \leq t_n, \quad \boldsymbol{\beta} = \left(\beta_{01}, \ldots, \beta_{0m}, \boldsymbol{\beta}_{\boldsymbol{x}}^T\right)^T$$

Joint Number of Reported Cases as Observed Responses of a Multinomial Model

Another popular approach is to consider that the observed responses arise from a Multinomial model, to apply models for such responses and to estimate the model parameters parameters using methods for missing data.

Let Y_{ij} be the random variable representing the number of HIV-AIDS cases diagnosed at the time unit t_j with reporting delay d_i. For a given j, the conditional distribution of $Y_j = \{Y_{0j}, \ldots, Y_{d_m j}\}^T$ at a given $Y_{+j} = \sum_{k=0}^{d_m} Y_{kj}$ is multinomial with the trials' probabilities given by $f_{0j}(\mathbf{Z_j}), \ldots, f_{d_m j}(\mathbf{Z_j})$, that is $Y_j \mid Y_{+j} \sim MN(Y_{+j}, f_{0j}(\mathbf{Z_j}), \ldots, f_{d_{mj}}(\mathbf{Z_j}))$.

Remember that, for a given t_j, only Y_{ij} satisfying $t_j + d_i \leq t_n$ are observed. So, Y_j can be re-written as:

$$Y_j = \begin{cases} Y_j^C = \left\{Y_{0j}^C, \ldots, Y_{d_m j}^C\right\}^T & \text{if } t_0 \leq t_j \leq t_n - d_m \\ Y_j^{Tr} = \left\{Y_{0j}^{Tr}, \ldots, Y_{d_{t_n - t_i} j}^{Tr}\right\}^T & \text{if } t_n - d_m + 1 \leq t_j \leq t_n \end{cases} \tag{4.83}$$

Note that, if $t_0 \leq t_j \leq t_n - d_m$, $Y_{+j} = Y_{+j}^C$ otherwise $Y_{+j} > Y_{+j}^{Tr}$. However , in each case the number of reported cases follows a multinomial distribution, with sample sizes Y_{+j}^C and Y_{+j}^{Tr} and probabilities

$$f_j = \begin{cases} f_j^C = \left\{f_{0j}^C(\mathbf{Z_j}), \ldots, f_{d_{mj}}^C(\mathbf{Z_j})\right\}^T & \text{if } t_0 \leq t_j \leq t_n - d_m \\ f_j^{Tr} = \left\{f_{0j}^{Tr}(\mathbf{Z_j}), \ldots, f_{d_{t_n - t_i} j}^{Tr}(\mathbf{Z_j})\right\}^T & \text{if } t_n - d_m + 1 \leq t_j \leq t_n \end{cases} \tag{4.84}$$

respectively.

A common model for the reporting delay distribution that arises from a multinomial response is:

$$f_{ij}(d_i \mid \mathbf{Z_j}) = \frac{\exp(\eta_{ij}(\mathbf{Z_j}))}{\sum_{d_i=0}^{d_{mj}} \exp(\eta_{ij}(\mathbf{Z_j}))} \tag{4.85}$$

where d_{mj} is the longest reporting delay that can be reliably estimated in t_j and $\eta_{ij}(Z_j)$ is the log-linear

$$\eta_{ij}(\mathbf{Z_j}) = \begin{cases} 0 & \text{if } d_i = 0 \\ \alpha_{ij} + \mathbf{Z_{ij}}^T \beta_{ij} & \text{if } 1 \leq d_i \leq d_{mj} \end{cases} \tag{4.86}$$

where α_i and β_i are parameter vectors. This model implies that the probability that a case with covariate $\mathbf{Z_j}$ is reported at time t_j is multiplied by a factor $\exp\left(\mathbf{Z_{ij}}^T \beta_{ij}\right)$. Note that by letting $\alpha_{0j} = \beta_{0j} = 0$, (4.86) can be expressed as $\eta_{ij} = \alpha_{ij} + \mathbf{Z_{ij}}^T \beta_{ij}$ ($0 \leq d_{ij} \leq d_{mj}$) . For a further simplification, we let $\theta = (\alpha, \beta)$ and denote the dependence of $f_{ij}(d_{ij} \mid \mathbf{Z_j})$ on θ by $f_{ij}(d_{ij} \mid \mathbf{Z_j}, \theta)$ [8].

The maximum likelihood functions, which are the basis of the inferences, for these two approaches are analogous and are based on the multinomial - Poisson transformation described in the next sub-section.

4.2.4 Likelihood for the Joint Model

Many count problems lead to multinomial distributions, either unconditional or conditional[212]. The associated likelihood if often difficult to maximize but it can be transformed into a simpler form associated with a Poisson distribution and with additional parameters, still yielding identical estimates and asymptotic variances. This method can be applied to complete or incomplete data.

Multinomial - Poisson transformation for complete data

Assume that

$$\mathbf{Y} = \{Y_1, \ldots, Y_j, \ldots, Y_J\}$$

follows a multinomial distribution with parameters proportional to $\exp(\beta_j)$. Let $(y_1, \ldots, y_j, \ldots, y_J)$ be a realization of $\mathbf{Y}$. The maximum likelihood function is

$$L_M(\{\beta_j\}) = \frac{(y_1 + \ldots + y_J)!}{y_1! \ldots y_J!} \prod_{j=1}^{J} \left(\frac{\exp(\beta_j)}{\sum_{i=1}^{J} \exp(\beta_i)} \right)^{y_j} \tag{4.87}$$

Baker in [212], proved that maximizing equation (4.87) over $\{\beta_j\}$ provide the same estimates as maximizing the augmented model

$$L_P(\phi, \{\beta_j\}) = \frac{1}{y_1! \ldots y_J!} \prod_{j=1}^{J} (\exp(\phi + \beta_j))^{y_i} \exp(-\exp(\phi + \beta_j)) \tag{4.88}$$

over ϕ and $\{\beta_j\}$, which is the maximum likelihood for a vector of independent Poisson random variables: $Y_j \sim P(\exp(\phi + \beta_j))$, for $j = 1, \ldots, J$.

We consider a more general case. Let $\mathbf{Y}_i = \{Y_{i1}, \ldots, Y_{ij}, \ldots\}$, for $i = 1, 2, \ldots, I$, and $j \in J_i$ denote a vector of random variables with realization $y_i = \{y_{i1}, \ldots, y_{ij}, \ldots\}$. The subscript i indexes levels of a categorical covariate or a cross-classification of categorical covariates. Assume that $\mathbf{Y}_i \sim$ multinomial with parameters $\left\{ \frac{g_{ij}(\beta)}{G_i(\beta)}, j \in J_i \right\}$ where $G_i = \sum_{j \in J_i} g_{ij}(\beta)$ and $\beta = \{\beta_1, \ldots, \beta_q\}$. The maximum likelihood is

$$L_M(\beta) = \frac{\left(\sum_{i=1}^{I} \sum_{j=1}^{J_i} y_{ij} \right)!}{\prod_{i=1}^{I} \prod_{j=1}^{J_i} y_{ij}!} \prod_{i=1}^{I} \prod_{j=1}^{J_i} \left(\frac{g_{ij}(\beta)}{G_i(\beta)} \right)^{y_{ij}} \tag{4.89}$$

There are four approaches to maximize (4.89):

- Newton-Raphson algorithm without any transformations or reparametrizations;
- modification of the EM algorithm for truncated categorical data;
- especially for discrete time survival data, using a newton-raphson algorithm after a reparametrization using reverse time hazard functions;
- using a Multinomial-Poisson transformation transformation.

Let $\phi = (\phi_1, \ldots, \phi_i, \ldots \phi_I)^T$. The Multinomial - Poisson transformation of (4.89) is

$$L_p(\phi, \beta) = \prod_{i=1}^{I} \prod_{j \in J_i} \{g_{ij}(\beta) \exp \phi_i\}^{y_{ij}} \exp\{-g_{ij}(\beta) \exp \phi_i\} \tag{4.90}$$

The derivative of the logarithm of (4.90) with respect to ϕ_i is $\sum_j y_{ij} - G_i(\beta) \exp \phi_i$. Setting this derivative to 0, solving for $\hat{\phi}_i(\beta)$ and substituting into equation (4.90) gives $L_P\left(\phi\hat{(\beta)}, \beta\right) \propto L_M(\beta)$ [212]. These two likelihood functions have identical maximum likelihood estimates for β and their asymptotic variances (based on the observed information matrix) and therefore only (4.90) can be used to draw inferences about β.

4.3 Summary

Generalized linear models (GLM) are methods that allow the investigation of the effects of explanatory variables on response variables of several types. It is assumed that the distribution of the response belongs to the exponential family. GLMs consist of three components: a probability distribution from the exponential family for the response Y; a linear predictor $\eta = \mathbf{X}\beta$; and a link function g such that $E(Y) = \mu = g^{-1}(\mathbf{X}\beta)$. The model parameters can be estimated by the principle of maximum likelihood. A common choice for checking the model assumptions is through the graphical inspection of the residuals.

A multilevel model incorporates fixed-effects parameters and random parameters, which are unobserved random variables. The data has a grouped, nested or hierarchical structure such as: repeated measures, longitudinal and/or multilevel or hierarchical. In this model it is considered that the variance-covariance matrix of the residuals is simply a multiple of the identity matrix. Parameter estimates can be obtained by maximum likelihood or restricted maximum likelihood, which minimizes the standard errors of the regression coefficients estimates.. To asses the precision of the parameter estimates, it is used the profile deviance with respect to each of the parameters and applying a signed root transformation to the statistic of the likelihood ratio test. These calculations provide a profile zeta function for each parameter used to construct likelihood-based confidence intervals for the parameters. Profile zeta plots allow us to visually assess the precision of individual parameters and evaluate the models assumptions. Prediction intervals from the conditional distribution of the random effects, given the observed data, allow us to assess the precision of the random effects[179].

Survival analysis is a sub-field of statistics concerning the study of times to a particularly event. Whenever the events do not occur during the study period, the corresponding times are said to be censored data. Traditional methods have been largely developed and in addition, many machine learning algorithms have also been adapted to effectively handle survival [171]. In particular, the parametric method presents an easy to interpret, efficient and accurate method when the survival times follow a particular distribution. These models can be built and evaluated similarly to the generalized linear models.

The K-nearest neighbour algorithm is an unsupervised learning algorithm which performs a space partition in a pre-user-defined number of clusters (K). Any new observation is compared to a training sample. This new case is then assigned to the class corresponding to the majority vote from its K nearest neighbours.

Longitudinal k-means is also an unsupervised learning algorithm. The aim of this method is to determine sets of homogeneous and separated groups of trajectories. A trajectory is a sequence of observations of an individual measured at different times. This is an algorithm belonging to the EM class which alternates between the mean estimation and its assigning.

Naive Bayesian Classifier is a supervised learning algorithm and probabilistic classifier which assumes that all features contribute equally and independently to the final decision [173]. It combines a probability model based on the Naive Bayes conditional probability with a decision rule. The decision rule is simply the assignment of the observation to the class with the highest probability.

Multilayer perceptron networks, in particular the feed-forward single hidden layer networks, with back - propagation training algorithms, are the most common networks. They are supervised learning methods and can be used for classification or regression tasks. They are effective in the analysis of complex data with non - linear trends and time - dependent covariates, and even high-order interactions [187].

Support vector machines are the most popular and efficient classification and regression methods currently available. They are supervised learning methods that apply simple linear methods in a high-dimensional feature space that is non-linearly related to the input space.

Usually, all attributes are employed and non-overlapping partitions are generated.

In supervised methods, one of the most common strategies for model validation is reusing efficiently the sample by performing cross - validation.

Over almost 40 years, researchers have been interested in obtaining valid estimates of recent incidence of reportable diseases which are, inevitable, subject to delays. Accurate and timely estimates are necessary for research evidence guiding Public Health decision-makers.

Most of the approaches addressing the reporting delay focus on the joint modelling of the reporting delay and incidence of HIV-AIDS. In this setting the two major approaches are the reverse time hazard and the discrete outcome models.

The reverse time hazards approach focus on the reporting delay distribution and allow the visualization of the hazard function in continuous time. Moreover, the analysis of the effect of the covariates is done in a natural fashion. The long delay is taken into consideration and so the entire delay distribution is analysed. The main weakness of this approach is that, in general, it implies data transformation on the time axis and the interpretation of the results is more difficult.

The discrete outcome models focus on incidence and accommodate right truncation without transforming the data. They can be implemented with standard statistical packages. Usually Poisson and Multinomial assumptions are involved, which may yield biased inferences if data fails to meet the posited distributional assumptions.

Recently, separate approaches are being considered to provide more flexibility and complexity to the reporting delay and incidence sub-models. Each model can even be fitted in different samples. Like the discrete outcome approach, it may need parametric assumptions.

Chapter 5

Methodology and Results

Circumstances do not make the man, they reveal him. - James Allen

In the presence of reporting delays, only a portion of diagnosed patients in the i^{th} time period are observed within a limited time interval. As a consequence, the number of observed cases generally underestimates the disease true incidence. Given the importance of an accurate and timely estimate of the HIV-AIDS incidence, several approaches have been studied.

Traditionally, reporting delays are divided into yearly quarters and considered as a discrete outcome. This approach is convenient when there are considerable uncertainties in the dates of interest; moreover, in many countries, the cases are reported in batches and so, grouping the time dates accommodates these fluctuations.

In section 5.1 it is presented a description of the main features of the Portuguese Epidemic and how it was surveyed. In section 5.2 the reporting delays are grouped into yearly quarters and considered as a discrete outcome. An incidence and delay joint model approach is described in section 5.2.1. In this setting, a stop Poisson model is described for capturing the underlying phenomena of underreporting which is presented in 5.2.1. A model adding stationarity is presented in section 5.2.2 and non-parametric approaches are presented in section 5.2.3, namely: feed-forward, naive bayesian classifier, k-nearest neighbour and support vector machines. In section 5.3 the reporting delay is considered a continuous outcome and the full context of the complex system is taking into consideration through a multilevel approach.

5.1 The HIV - AIDS Portuguese Epidemic

Since the beginning of the epidemic, most of the reported cases were from asymptomatic cases, closely followed by AIDS cases. Following the observed pattern in the European countries, the majority of diagnosed cases were men. The heterosexual were the most common risk-group category in the absolute sense. In the relative sense, taking into consideration the size of the risk groups, there were a large amount of cases within injecting drug users (IDUs). The first case being diagnosed occurred in 1983 and the first notification was in 1985.

5.1.1 Incidence

Looking at the Portuguese trend of the number of cases by year of diagnosis (Figure 5.2), the epidemic suffered an increase in the year 2000, then progressively decreased until

around the year 2005, at which the number of reported and diagnosed cases became approximately constant until 2008; from then onwards, it started to descend again. It is important to bear in mind that the surveillance system is affected by reporting delays and so the number of cases in the latest years of the epidemic in this study most probably does not corresponds to the true incidence.

When the cases are analysed by year of notification, it can be seen an increasing tendency on the number of cases until the year 2000, a rapid decrease in the next year and, from there onwards, the number of notified cases has became roughly constant (Figure 5.1).

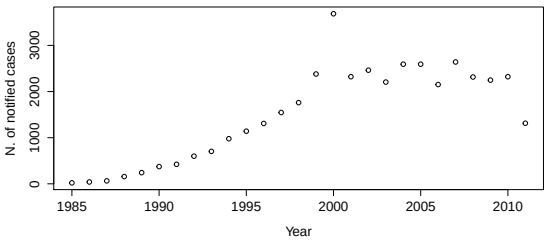


Figure 5.1: Number of cases by notification year

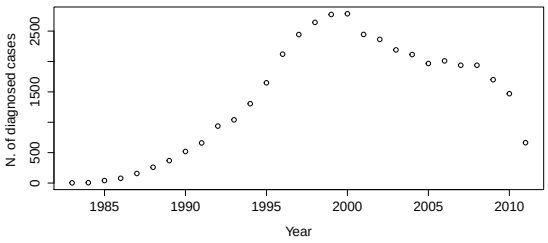


Figure 5.2: Number of cases by diagnosis year

Late diagnosis, that is AIDS cases, was the most reported stage in the early beginning of the epidemic and of the surveillance system. Around the year 2000, asymptomatic cases were reported more frequently being followed closely by the AIDS cases (Figure 5.3). The same pattern is observed when this characteristic is represented by the notification year (Figure 5.4).

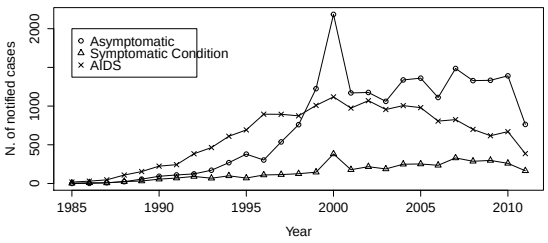


Figure 5.3: Number of AIDS cases per year of notification for each disease stage

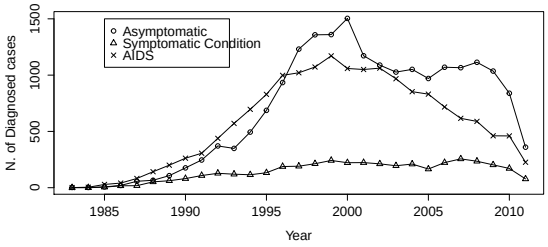


Figure 5.4: Number of AIDS cases per year of diagnosis for each disease stage

In the whole epidemic history, male cases have always been more frequently observed (Figures 5.5 and 5.6). Injection drug users (IDU's) were the most frequent risk group diagnosed and reported until around the year 2002; from that onwards, it gave place to the heterosexual risk group (Figures 5.7 and 5.8). The median age of infected individuals was around 40 years old, in the whole epidemic history considered in this study (Figures 5.9 and 5.10).

Throughout the years there have been several fluctuations on the number of cases per year of notification and region of health administration RHA's, (Figure 5.11). Nevertheless, Lisbon and the Tagus Valley and the North have been the regions with the majority of the contributions, most probably because they serve the two major Portuguese cities. It is important to notice a large contribution of cases coming from a non identified region.

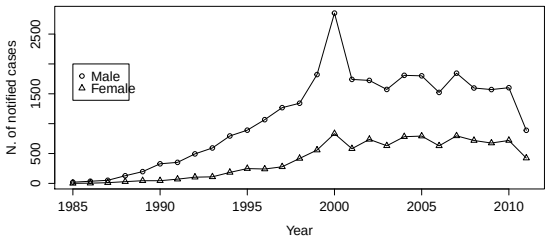


Figure 5.5: Number of notified cases per year for each sex

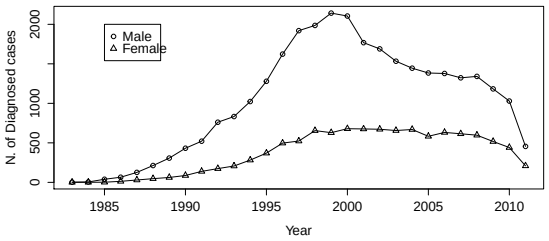


Figure 5.6: Number of diagnosed cases per year for each sex

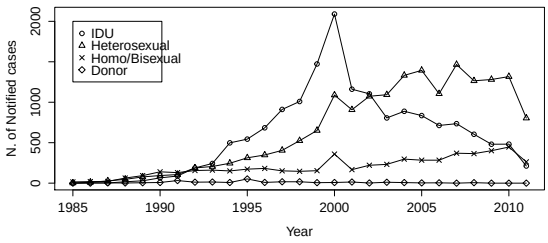


Figure 5.7: Number of notified cases per year for each risk group

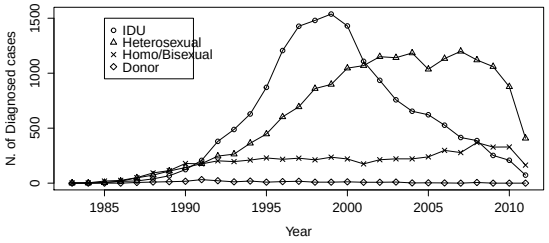


Figure 5.8: Number of diagnosed cases per year for each risk group

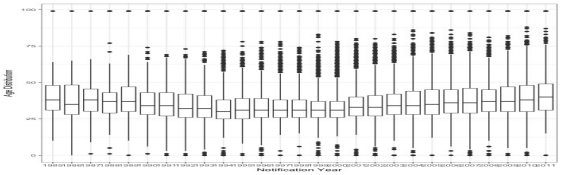


Figure 5.9: Age distribution per notification year

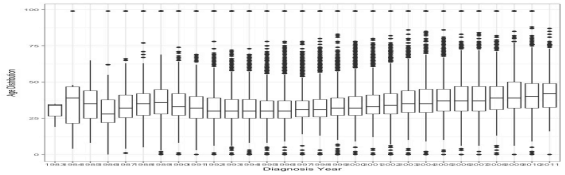


Figure 5.10: Age distribution per diagnosis year

From the perspective of the diagnosis year, the epidemic curves are smoother and naturally, the main contributions arise from the same regions as before (Figure 5.12) .

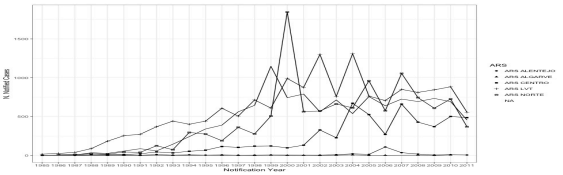


Figure 5.11: Number of HIV - AIDS cases per RHA and per notification year

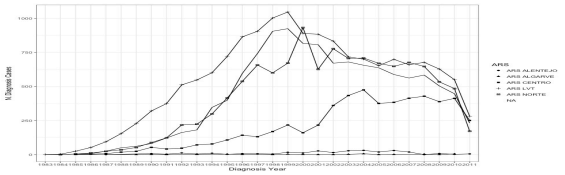


Figure 5.12: Number of HIV - AIDS cases per RHA and per diagnosis year

Considering all stages, the observed number of diagnosed HIV-AIDS cases in Portugal, from 1983 to 2011, is presented in Figure 5.13. This indicator exhibited an increasing trend between 1983 and the year 2000, four years after the introduction of HAART. Since then, the number of diagnosed cases has been steadily decreasing. When the notification became mandatory, a slight growth was observed.In the lower panel of the figure, we have included information about the historical events that may have introduced bias in the preva-

lence numbers. It can be seen that the prevalence curve of AIDS changes slightly at those moments.

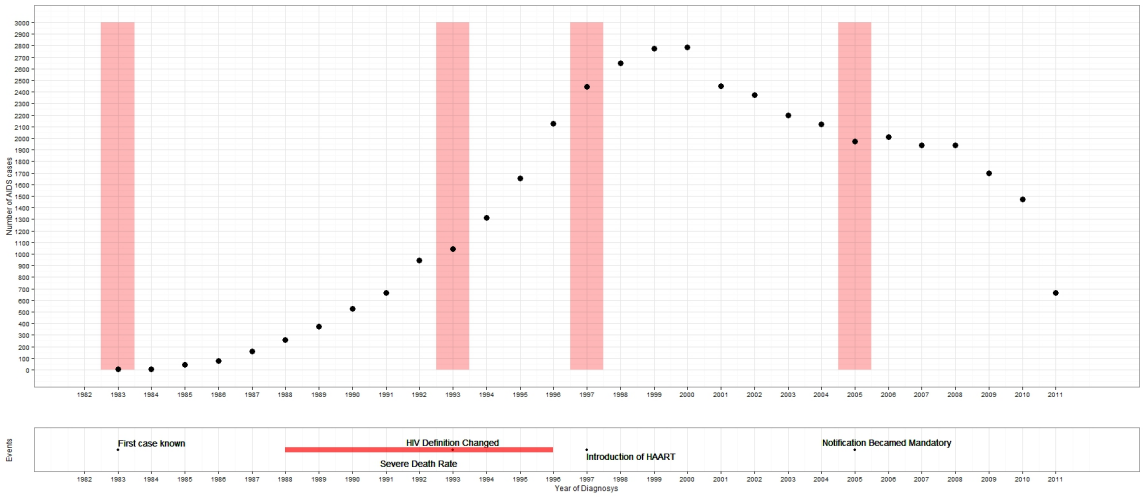


Figure 5.13: Number of AIDS diagnosed per year and epidemiological events

Death related to AIDS cases is reported through a notification form to CVEDT and to Instituto Nacional de Estatística (INE) based on death certificates. The difference between the two systems is represented in Figures 5.14, 5.15 and 5.16. It is clear that the underreporting of death within the HIV -AIDS surveillance system is high; since this is a very special case, it cannot be used for estimating the underreporting behaviour at other stages.

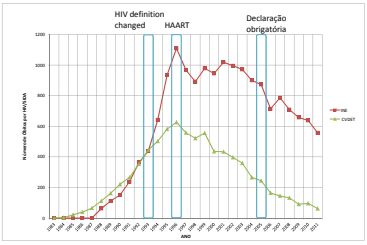


Figure 5.14: Mortality by INE and CVEDT

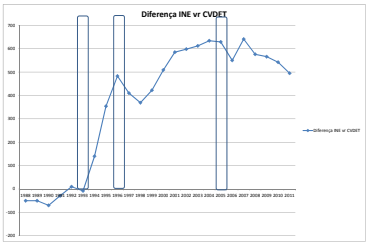


Figure 5.15: Difference from INE and CVEDT

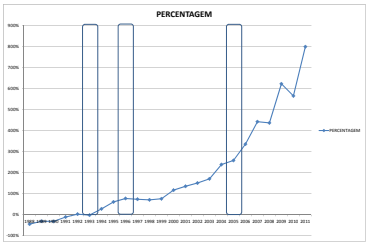


Figure 5.16: Mortality difference percentage INE and CVEDT

5.1.2 Reporting Delay

Considering the importance of the reporting delay on the timely ascertainment of the cases, we focus our attention on that issue in this section.

Moreover, it is less likely to have underreporting at the last stage of the disease due to the severe symptoms and the need for care so, in a initial analysis, we will restrict ourselves to the AIDS cases.

Considering the notification year, most of the cases have been reported between 1994 and 2006 with a reporting delay up to two years. The same pattern has been observed when considering the diagnosis year. There is a considerable long reporting delay through all years of notification and diagnosis; for example, in the year 2000 approximately 200

cases had a reporting delay of 4 years (Figures 5.17 and 5.18). There is an increasing pattern in the number of cases until the year 2000 and a decreasing pattern afterwards, which resembles the incidence pattern represented in (Figure 5.13).

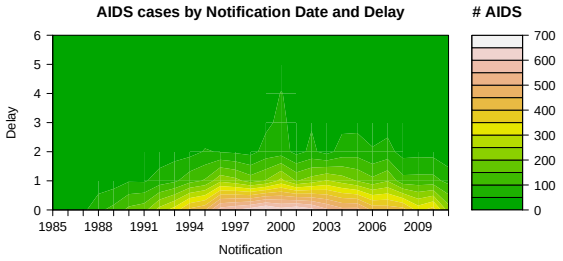


Figure 5.17: Number of cases of AIDS per Notification year and reporting delay by trimester

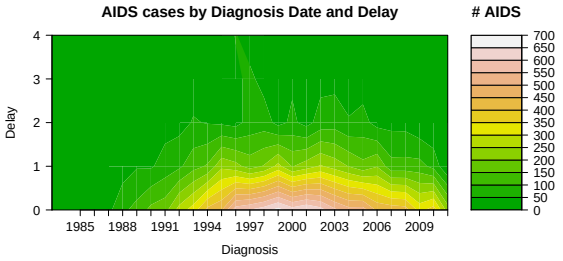


Figure 5.18: Number of cases of AIDS per Diagnosis year and reporting delay by trimester

Taking into consideration the annual incidence, the percentage of AIDS cases (within all cases) by reporting delay and year of diagnosis is presented in Figure 5.19. From this representation, it can be seen that the majority of cases are reported within one year.

To get a better description of the patterns of the observed delays, the time intervals (diagnosis, notification, delay,...) were divided into trimesters discarding the first three years of the disease, since in the early days there were not enough knowledge and stability of the surveillance system for the true pattern to be accurately captured.

Figure 5.20 depicts the annual percentage of cases within each diagnosis year and reporting delay. The interpretation of the curve in the most recent years (shaded region) must be done carefully since there are cases that have not been notified yet.

Most of the cases are reported within 3 months after diagnosis but some of them are still being reported with a delay longer than one year. For the sake of clarity, delays longer than 18 months are omitted from Figure 5.20 (the longer delay occurs with a low frequency).

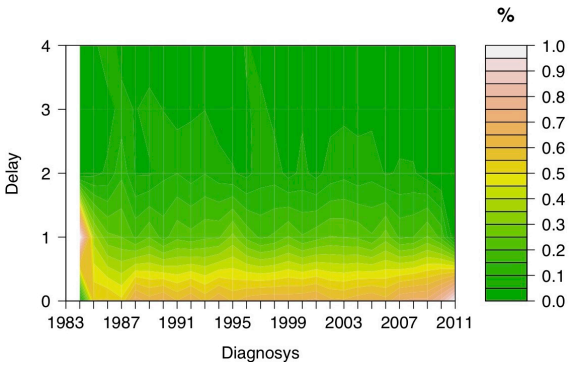


Figure 5.19: Percentage of AIDS cases per diagnosis year and reporting delay year

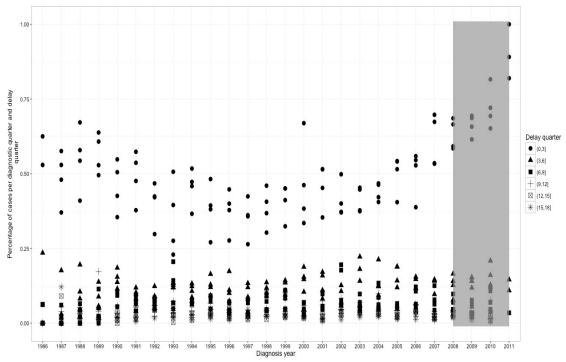


Figure 5.20: Percentage of HIV-AIDS cases per diagnosis and delay quarters. The lighted region identifies those (recent) years that have to be corrected.

Looking closely at the delay curves, the percentage (considering all cases diagnosed in the trimester) of cases reported within 3 months has been increasing. That means that, in the most recent years,the system is registering cases with a more reasonable delay time and the percentage of cases reported within 3 to 6 months has been relatively stable

between 10% and 20% (Figure 5.21).

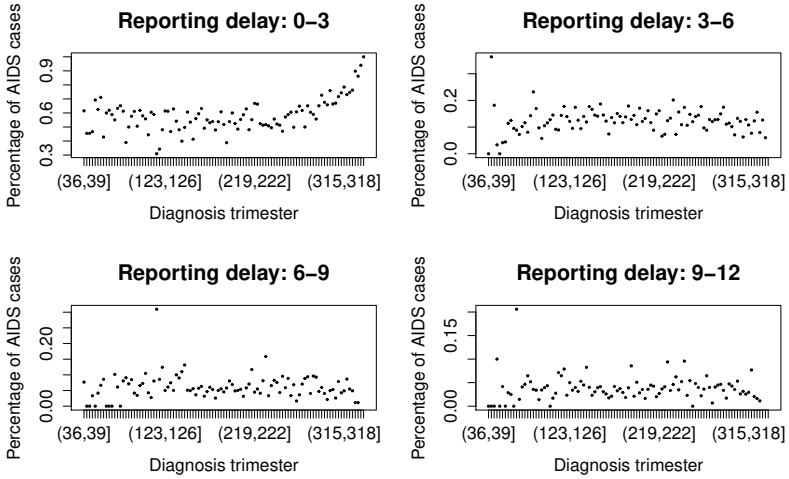


Figure 5.21: Separate delay curves for percentage of HIV-AIDS cases per diagnosis and delay quarters. The lighted region identifies those (recent) years that have to be corrected.

The longitudinal distribution of the reporting delays according to several demographical and institutional variables are described in Figures 5.22 to 5.27. In these pictures, the delays are divided into 3 months long intervals and are identified as: 0 representing the quarter $[0, 3]$ months , 1 the quarter $[3, 6]$ months, 2 the quarter $[6, 9]$ months and 3 the longer delays.

The infection is concentrated at the male gender (Figure 5.22) and most of the cases have, naturally, an European nationality being followed by Africans (Figure 5.23). From these representations, it is clear a large concentration of cases with delay $[0, 3]$ months; in the case of the European citizens, there is also a considerable amount of cases with other reporting delay lengths.

Early detections (corresponding to asymptomatic cases) were most of the reported cases, especially in the most recent years. These cases presented equally distributed reporting delays through the quarters (Figure 5.24). Large concentrations of AIDS - cases (late detections and/or disease progressions) must also be noticed, mostly with a reporting delay lower than 3 months.

As for the risk-groups, the infection has been most prevalent amongst the heterosexual community, mainly after 1999, which is natural due to the large size of this group. A large frequency of IDUs is also visible from 1996 to 2003 (Figure 5.25). The cases are approximately equally distributed by the different reporting delays and diagnosis years.

The major reporting institutions are the hospitals, and most of the reporting delays are lower than 3 months (Figure 5.26).

Considering the regional health administration, it can be seen that RHA North had the majority of the reported cases. It can also be noticed that the time lag between diagnosis and reporting is, in most cases, less than 3 months (Figure 5.27).

Considering the reporting delay grouped into only two groups - the $0-3$ months and > 3 months - the distribution of the number of cases according to sex, age, nationality, disease stage, HIV risk-group, type of health care institution and regional health administration is presented in Table 5.1. The results show that the two groups have similar characteristics with the exception of “Disease Stage” and “HIV Risk Group”. The majority of the cases in the faster group are AIDS cases and in the slower group are asymptomatic cases. This may be due to the urge of treatment for these cases. Concerning the risk group, a slight difference is observed for the two groups: in the faster group, heterosexuals are the prevalent risk-group followed by the IDU’s; in the slower group, this pattern is still visible but the difference

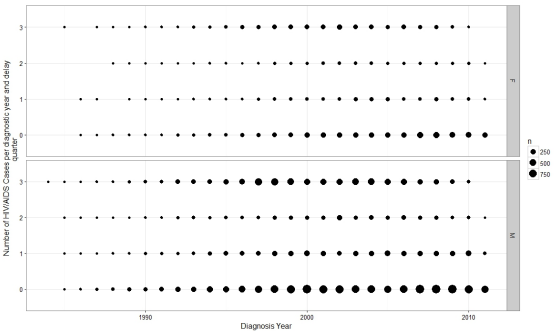


Figure 5.22: Number of HIV-AIDS infections by year of diagnosis (x-axis), reporting delay quarter (y-axis) and gender (right vertical boxes)

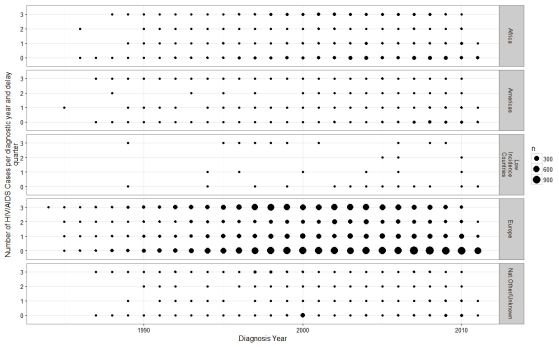


Figure 5.23: Number of HIV-AIDS infections by year of diagnosis (x-axis), reporting delay quarter (y-axis) and patients nationality (right vertical boxes)

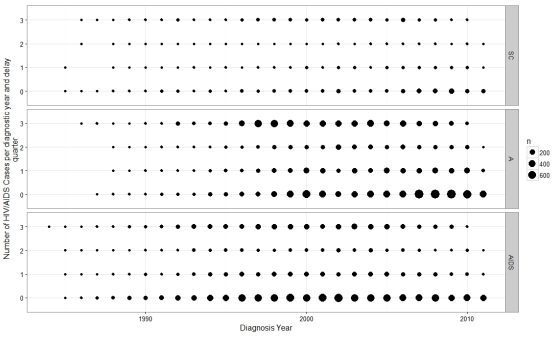


Figure 5.24: Number of HIV-AIDS infections by year of diagnosis (x-axis), reporting delay quarter (y-axis) and disease status (right vertical boxes).

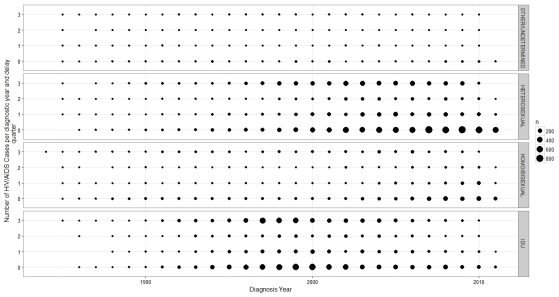


Figure 5.25: Number of HIV-AIDS infections by year of diagnosis (x-axis), reporting delay quarter (y-axis) and risk group (right vertical boxes)

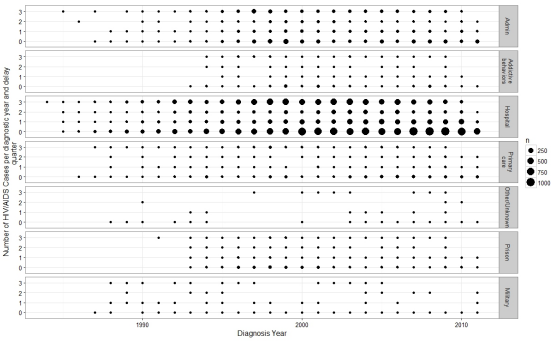


Figure 5.26: Number of HIV-AIDS infections by year of diagnosis (x-axis), reporting delay quarter (y-axis) and health care institution (right vertical boxes)

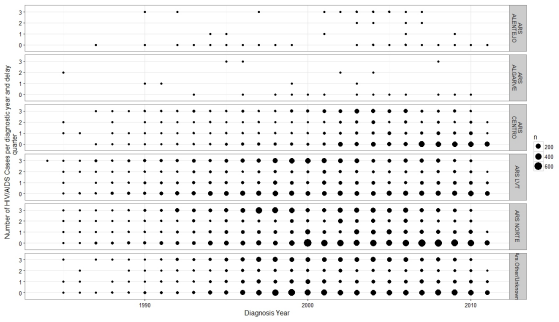


Figure 5.27: Number of HIV-AIDS infections by year of diagnosis (x-axis), reporting delay quarter (y-axis) and regional administration (right vertical boxes)

is narrower.

Table 5.1: Overview the two classes of reporting delay per demographic variables

Variable		0-3		>3	
		mean	std	mean	std
Age at diagnosis		37.74	13.03	36.21	12.93
		n	%	n	%
Gender	Female	4397	24	4832	26
	Male	13815	76	13466	74
Patient Nationality	Africa	1786	41	1599	53
	Americas	448	2	361	2
	Europe	15309	84	15688	86
	Low Incidence Countries	30	0	23	0
	Nat. Other - Unknown	639	4	627	3
Disease stage	A	7501	41	9777	53
	AIDS	9063	50	6712	37
	SC	1648	9	1809	10
HIV Risk-group	Heterosexual	8491	47	7683	42
	Homo - Bisexual	2778	15	2575	14
	IDU	6331	35	7452	41
	Other - Undetermined	612	3	588	3
Health Care Institution	Centres for treatment of addictive behaviours	457	3	296	2
	Unspecific Administration	2346	13	2898	16
	Hospital	13653	75	13913	76
	Primary care	1033	6	716	4
	Prisons	555	3	358	2
	Military	103	1	80	0
	Other - Unknown	65	0	37	0
Regional Health Administration	Alentejo	91	0	109	1
	Algarve	28	0	10	0
	Centro	2367	13	2222	12
	Lisboa e Vale do Tejo	4103	23	4742	26
	Norte	6581	36	6381	35
	Other - Unknown	5041	28	4834	26

5.2 Discrete Outcomes

5.2.1 Joint Modelling - Count Regression Models

The presence of this section in the thesis is due to the popularity of the joint modelling and the fact that the methodology has already been applied to the Portuguese data by others [203, 204].

Consider the cases cross-classified by the notification and diagnosis dates and focus only in the most 'visible' stage of the disease - the AIDS stage. The number of AIDS cases cross-classified by the notification year and the diagnosis year is presented in Table 5.2. Some important features of that table are as follows:

1. Cells in the upper region are necessarily empty because of the logical relationship between the diagnosis and the notification year;
2. Most of the cases are at the main diagonal and at the main rows immediately below it;
3. The cases at the final columns are partially observed due to the presence of reporting delay.

As the two dimensions of this table are the year of diagnosis and reporting delay, the

margins for the diagnosis year give the total incidence over time, but with the most recent values masked by the reporting delay.

		Year of Diagnosis																													
		1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	Sum
Year of Notification	1985	0	2	16																											18
	1986	0	0	11	19																										30
	1987	0	0	1	9	37																									47
	1988	0	0	0	2	18	89																								109
	1989	1	0	1	1	10	25	116																							154
	1990	0	0	0	3	5	6	46	164																						224
	1991	0	1	0	1	2	6	8	44	181																					243
	1992	0	1	0	1	2	4	13	22	71	270																				384
	1993	0	0	0	0	2	4	2	12	30	90	324																			464
	1994	0	0	0	4	2	6	6	4	8	32	132	416																		610
	1995	0	0	0	0	0	0	2	4	11	18	54	156	447																	692
	1996	0	0	0	0	0	0	0	1	0	4	27	47	238	578																895
	1997	0	0	0	0	0	1	2	1	1	2	13	14	41	217	602															894
	1998	0	0	0	0	0	0	0	1	0	5	2	9	16	33	178	630														874
	1999	0	0	0	0	1	0	3	0	1	2	1	8	12	40	68	204	669													1009
	2000	0	0	0	0	0	1	1	2	1	1	5	11	17	53	61	74	266	626												1119
	2001	0	0	0	1	2	0	0	0	0	0	0	8	8	14	26	32	34	197	652											974
	2002	0	0	0	0	0	0	0	0	0	2	3	8	18	27	24	35	46	60	225	622										1070
	2003	0	0	0	0	0	0	0	1	0	1	1	0	1	5	10	16	29	41	33	256	562									956
	2004	0	0	0	0	0	0	0	1	1	4	3	6	12	12	12	24	40	36	37	70	228	520								1006
	2005	0	0	0	0	0	0	0	0	0	1	2	4	4	8	16	17	26	37	31	35	77	198	524							980
	2006	0	0	0	0	0	0	1	1	0	2	2	3	5	0	6	8	16	16	26	26	35	53	165	441						806
	2007	0	0	0	0	0	0	0	0	0	1	0	0	4	3	5	11	9	18	10	15	19	32	67	197	434					825
	2008	0	0	0	0	0	0	0	0	0	1	0	4	1	3	4	2	9	10	9	15	25	17	26	31	117	425				699
	2009	0	0	0	0	1	0	0	2	0	1	1	1	2	3	8	14	17	8	15	11	10	9	16	14	34	118	332			617
	2010	0	0	0	0	0	0	0	1	0	1	1	1	3	0	1	2	9	9	9	10	11	18	19	28	20	33	118	376		670
	2011	0	0	0	0	0	0	0	1	1	0	0	0	0	2	0	3	1	1	3	2	1	6	14	6	11	13	11	84	226	386
	Sum	1	4	29	41	82	142	200	262	306	438	571	696	829	998	1021	1072	1171	1059	1050	1062	968	853	831	717	616	589	461	460	226	16755

Table 5.2: Number of cases per notification and diagnosis year

The number of cases cross-classified by reporting delay and diagnosis year is represented in Figure 5.29 and the number of cases taking into consideration the reporting delay and the notification year is presented in Figure 5.28. Both approaches include the marginal distributions. From these representations, it is clear the long tailed distribution of the reporting delay as well as its extreme skewness. The marginal distribution of the notification year and diagnosis year have already been discussed in sub-section 5.1.1.

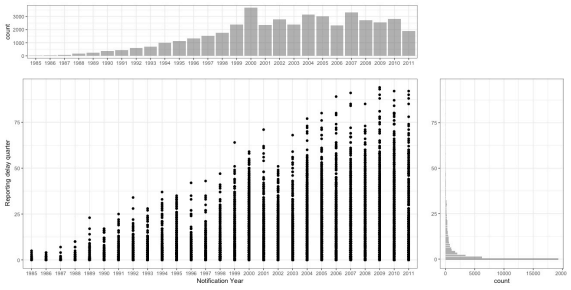


Figure 5.28: Number of cases per notification year and reporting delay with marginal distributions of reporting delay and number of cases per notification year

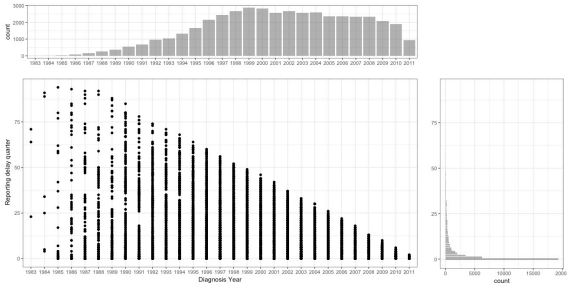


Figure 5.29: Number of cases per diagnosis year and reporting delay with marginal distributions for the reporting delay and incidence

Again, we focus at AIDS cases due to their lower probability of underreporting. The above table displays how AIDS cases are occurring in the population over time and how these cases, occurring at a given time point, subsequently arrive at the central office [176]. The latter process may be evolving over time, for example, if reporting improves or if the increasing number of cases swamps available facilities [176].

This bivariate process may be formulated as a bivariate count process, which can be fitted as:

$$X_{ij} = \gamma + \alpha Delay_i + \beta Diagnosis_j \tag{5.1}$$

where X_{ij} is the number of cases cross-classified by their delay length i and time of diagnosis j, with $Delay$ and $Diagnosis$ divided into time intervals and appropriately considered as categorical variables.

Let T_0 be the date of the first reported diagnosed case and T_n be the most recent date. Consider that the time interval $[T_0, T_n]$ of diagnosis is divided into equal length intervals, say $[T_{j-1}, T_j]$ with $j = 1, \ldots, n$. Let d_0 be the shorter reporting delay observed and d_m be the longest. Consider that the reporting delay interval $[d_0; d_m]$ is also divided into equal length intervals, say $[d_{(i-1)}; d_i]$ with $i = 1, \ldots, m$. Let X_{ij} be the random variable representing the number of AIDS cases diagnosed in the j^{th} calendar interval $[T_{j-1}, T_j]$ whose reporting delay falls on the interval $[d_{i-1}, d_i]$. Note that only AIDS cases whose diagnosis date plus the reporting delay is less than T_n are recorded [203, 204]. Thus, only X_{ij} satisfying $1 \leq i \leq m_j$, where $m_j = \max\{1, \ldots, m$ such that $T_j + d_{m_j} \leq T_n\}$ are accessible. Whenever $T_j + d_{m_j} = T_n$, the variable X_{ij} is only partially observed.

The main statistical purpose in estimating the reporting delay distribution is to draw inferences about an unknown vector θ of parameters that characterizes the joint distribution of count random variables whose realizations are right-truncated [121].

The estimation of the reporting delay distribution has been based on Poisson regression models assuming a stationary reporting delay process and estimation by conditional likelihood [203, 204, 199, 208, 121]. However, since Table 5.2 presents many zeros and too much variation in the magnitude of the numbers, besides the fitting of the classical Poisson process other models have also been considered: over-dispersed Poisson, Negative Binomial, Zero-Inflated Poisson, Zero-Inflated Negative Binomial, Hurdle Poisson and Hurdle Negative Binomial. The fit from these models is presented in Table 5.3.

The time intervals consisted of yearly quarters, a maximum delay of 9.5 years was assumed (the other cases being considered as noise) and the first 39 months of registries were included (from 1983 until 1986), due to the lack of quality in the data.

Comparing the estimates produced by all models, it can be seen that they do not vary much. Also, the estimates of the effects attributed to the reporting delay decrease as the length of the delay increases but the standard deviation is fairly constant until the delays get larger than 96 months (approximately 8 years). In terms of the effect of the diagnosis, the estimated effect increases until approximately month 120 (corresponding to the year 1993 when HAART was introduced); thereafter, the estimates slightly oscillated between 2.29 and 3.29 until the trim 288 (corresponding to the year 2007) where it started to slowly decrease.

Table 5.3: Count regression models using as predictors the delays quarter and diagnosis quarter

	Poisson	**Poisson overdispersion**	**Negative Binomial**	**Zero-Inflated Poisson**	**Zero-Inflated Negative Binomial**	**Hurdle Poisson**	**Hurdel Negative Binomial**
	value (sd)	value (sd)	value (sd)	value (sd)	value (sd)	value (sd)	value (sd)
(Intercept)	1.96 (0.28)	1.96 (0.35)	2.00 (0.33)	2.01 (0.35)	1.97 (0.29)	1.96 (0.35)	1.92 (0.42)
Delay (3,6]	-1.48 (0.02)	-1.48 (0.03)	-1.55 (0.06)	-1.48 (0.02)	-1.54 (0.052)	-1.48 (0.02)	-1.53 (0.05)
Delay (6,9]	-2.18 (0.03)	-2.18 (0.04)	-2.24 (0.06)	-2.17 (0.03)	-2.23 (0.06)	-2.17 (0.03)	-2.21 (0.05)
Delay (9,12]	-2.65 (0.04)	-2.65 (0.05)	-2.74 (0.07)	-2.63 (0.04)	-2.71 (0.06)	-2.64 (0.04)	-2.7 (0.06)
Delay (12,15]	-2.92 (0.05)	-2.92 (0.06)	-3.00 (0.07)	-2.88 (0.047)	-2.96 (0.07)	-2.88 (0.047)	-2.95 (0.06)
Delay (15,18]	-3.26 (0.06)	-3.26 (0.07)	-3.34 (0.08)	-3.21 (0.06)	-3.28 (0.07)	-3.22 (0.06)	-3.27 (0.07)
Delay (18,21]	-3.43 (0.06)	-3.43 (0.07)	-3.53 (0.08)	-3.34 (0.06)	-3.45 (0.08)	-3.35 (0.06)	-3.43 (0.08)
Delay (21,24]	-3.68 (0.07)	-3.68 (0.08)	-3.77 (0.09)	-3.59 (0.07)	-3.70 (0.09)	-3.63 (0.08)	-3.72 (0.09)
Delay (24,27]	-3.75 (0.07)	-3.75 (0.09)	-3.85 (0.09)	-3.66 (0.07)	-3.76 (0.09)	-3.65 (0.08)	-3.73 (0.09)
Delay (27,30]	-3.93 (0.08)	-3.93 (0.10)	-4.02 (0.10)	-3.77 (0.09)	-3.89 (0.10)	-3.73 (0.09)	-3.80 (0.10)
Delay (30,33]	-3.97 (0.08)	-3.97 (0.10)	-4.08 (0.10)	-3.87 (0.08)	-4.00 (0.10)	-3.88 (0.10)	-3.97 (0.11)
Delay (33,36]	-4.04 (0.08)	-4.04 (0.10)	-4.16 (0.10)	-3.89 (0.09)	-4.04 (0.10)	-3.89 (0.10)	-4.00 (0.11)
Delay (36,39]	-4.17 (0.09)	-4.17 (0.11)	-4.30 (0.11)	-4.06 (0.09)	-4.20 (0.11)	-4.07 (0.11)	-4.19 (0.12)
Delay (39,42]	-4.39 (0.10)	-4.39 (0.12)	-4.50 (0.11)	-4.25 (0.10)	-4.39 (0.12)	-4.18 (0.12)	-4.29 (0.14)
Delay (42,45]	-4.24 (0.09)	-4.24 (0.12)	-4.37 (0.11)	-4.14 (0.10)	-4.29 (0.11)	-4.13 (0.11)	-4.23 (0.13)
Delay (45,48]	-4.57 (0.11)	-4.57 (0.14)	-4.70 (0.12)	-4.46 (0.11)	-4.61 (0.13)	-4.34 (0.14)	-4.47 (0.16)
Delay (48,51]	-4.41 (0.10)	-4.41 (0.13)	-4.53 (0.12)	-4.32 (0.11)	-4.44 (0.12)	-4.34 (0.13)	-4.44 (0.15)
Delay (51,54]	-4.64 (0.11)	-4.64 (0.14)	-4.78 (0.13)	-4.49 (0.12)	-4.68 (0.13)	-4.28 (0.15)	-4.43 (0.16)
Delay (54,57]	-4.6 (0.11)	-4.62 (0.14)	-4.77 (0.13)	-4.48 (0.12)	-4.65 (0.13)	-4.20 (0.14)	-4.33 (0.16)
Delay (57,60]	-4.53 (0.11)	-4.5 (0.14)	-4.68 (0.13)	-4.40 (0.12)	-4.58 (0.13)	-4.26 (0.14)	-4.38 (0.15)
Delay (60,63]	-4.8 (0.13)	-4.8 (0.16)	-4.96 (0.14)	-4.72 (0.13)	-4.87 (0.14)	-4.37 (0.16)	-4.51 (0.18)
Delay (63,66]	-4.72 (0.12)	-4.72 (0.15)	-4.85 (0.14)	-4.62 (0.13)	-4.77 (0.14)	-4.37 (0.16)	-4.48 (0.18)
Delay (66,69]	-4.97 (0.14)	-4.97 (0.17)	-5.12 (0.15)	-4.86 (0.14)	-5.03 (0.15)	-4.68 (0.20)	-4.84 (0.21)
Delay (69,72]	-4.80 (0.13)	-4.80 (0.16)	-4.94 (0.14)	-4.66 (0.14)	-4.84 (0.14)	-4.51 (0.18)	-4.66 (0.19)
Delay (72,75]	-4.94 (0.14)	-4.94 (0.17)	-5.10 (0.15)	-4.81 (0.15)	-4.98 (0.15)	-4.44 (0.18)	-4.59 (0.20)
Delay (75,78]	-5.33 (0.17)	-5.33 (0.21)	-5.47 (0.18)	-5.22 (0.17)	-5.37 (0.18)	-5.29 (0.32)	-5.43 (0.33)
Delay (78,81]	-5.33 (0.17)	-5.33 (0.21)	-5.47 (0.18)	-5.23 (0.18)	-5.38 (0.18)	-5.28 (0.32)	-5.44 (0.33)
Delay (81,84]	-5.13 (0.16)	-5.13 (0.19)	-5.28 (0.17)	-5.02 (0.16)	-5.18 (0.17)	-4.82 (0.23)	-4.97 (0.25)
Delay (84,87]	-5.25 (0.17)	-5.25 (0.21)	-5.39 (0.18)	-5.16 (0.17)	-5.30 (0.18)	-4.87 (0.26)	-5.01 (0.27)

(Table continues...)

Table 5.3: Count regression models using as predictors the delays quarter and diagnosis quarters

	Poisson	Poisson Overdispersion	Negative Binomial	Zero-Inflated Poisson	Zero-Inflated Negative Binomial	Hurdle Poisson	Hurdle Negative Binomial
	value (sd)	value (sd)	value (sd)	value (sd)	value (sd)	value (sd)	value (sd)
Delay (87,90]	-5.56 (0.20)	-5.56 (0.24)	-5.70 (0.21)	-5.43 (0.20)	-5.60 (0.21)	-4.9 (0.31)	-5.09 (0.33)
Delay (90,93]	-5.27 (0.17)	-5.27 (0.21)	-5.41 (0.18)	-5.11 (0.18)	-5.30 (0.19)	-4.24 (0.21)	-4.39 (0.23)
Delay (93,96]	-5.44 (0.19)	-5.44 (0.24)	-5.59 (0.20)	-5.33 (0.19)	-5.49 (0.20)	-5.24 (0.34)	-5.41 (0.35)
Delay (96,99]	-5.98 (0.25)	-5.98 (0.31)	-6.13 (0.26)	-5.87 (0.25)	-6.04 (0.26)	-5.51 (0.48)	-5.71 (0.49)
Delay (99,102]	-6.02 (0.26)	-6.02 (0.32)	-6.17 (0.27)	-5.91 (0.26)	-6.07 (0.27)	-5.26 (0.48)	-5.44 (0.49)
Delay (102,105]	-5.34 (0.19)	-5.34 (0.23)	-5.50 (0.20)	-5.19 (0.19)	-5.39 (0.2)	-4.30 (0.23)	-4.51 (0.25)
Delay (105,108]	-6.29 (0.31)	-6.29 (0.38)	-6.42 (0.31)	-6.18 (0.31)	-6.33 (0.31)	-21.33 (1775.63)	-18.38 (361.69)
Delay (108,111]	-5.83 (0.24)	-5.83 (0.30)	-5.98 (0.25)	-5.72 (0.25)	-5.89 (0.25)	-5.31 (0.43)	-5.50 (0.44)
Delay (111,114]	-6.34 (0.32)	-6.34 (0.39)	-6.49 (0.32)	-6.22 (0.32)	-6.39 (0.32)	-6.44 (0.98)	-6.71 (0.99)
Diagnosis (39,42]	-0.17 (0.41)	-0.17 (0.51)	-0.13 (0.48)	-0.21 (0.51)	-0.13 (0.44)	-0.03 (0.50)	0.10 (0.60)
Diagnosis (42,45]	-0.17 (0.41)	-0.17 (0.51)	-0.09 (0.48)	-0.24 (0.46)	-0.11 (0.45)	-0.33 (0.54)	-0.28 (0.64)
Diagnosis (45,48]	0.84 (0.33)	0.84 (0.41)	1.01 (0.41)	0.77 (0.39)	0.97 (0.37)	0.66 (0.42)	0.74 (0.51)
Diagnosis (48,51]	0 (0.39)	0 (0.49)	-0.04 (0.48)	0.18 (0.48)	0.22 (0.53)	0.19 (0.49)	0.21 (0.59)
Diagnosis (51,54]	0.61 (0.34)	0.61 (0.43)	0.59 (0.43)	0.54 (0.40)	0.59 (0.38)	0.58 (0.42)	0.57 (0.52)
Diagnosis (54,57]	1.24 (0.31)	1.24 (0.40)	1.10 (0.40)	1.21 (0.38)	1.11 (0.35)	1.30 (0.39)	1.22 (0.48)
Diagnosis (57,60]	0.99 (0.32)	0.99 (0.40)	1.18 (0.40)	0.92 (0.38)	1.14 (0.36)	0.80 (0.41)	0.89 (0.50)
Diagnosis (60,63]	1.12 (0.32)	1.12 (0.40)	1.11 (0.40)	1.10 (0.39)	1.10 (0.35)	1.10 (0.40)	1.08 (0.49)
Diagnosis (63,66]	0.48 (0.35)	0.48 (0.44)	0.47 (0.44)	0.47 (0.43)	0.48 (0.46)	0.55 (0.43)	0.59 (0.54)
Diagnosis (66,69]	0.96 (0.33)	0.96 (0.41)	0.95 (0.41)	1.05 (0.39)	1.18 (0.40)	1.12 (0.40)	1.28 (0.50)
Diagnosis (69,72]	1.67 (0.30)	1.67 (0.38)	1.75 (0.38)	1.60 (0.37)	1.72 (0.34)	1.66 (0.38)	1.76 (0.47)
Diagnosis (72,75]	1.33 (0.31)	1.33 (0.39)	1.25 (0.40)	1.27 (0.38)	1.25 (0.35)	1.38 (0.39)	1.39 (0.48)
Diagnosis (75,78]	1.20 (0.32)	1.20 (0.40)	1.11 (0.40)	1.22 (0.38)	1.21 (0.38)	1.28 (0.39)	1.26 (0.49)
Diagnosis (78,81]	1.56 (0.31)	1.56 (0.38)	1.51 (0.39)	1.52 (0.38)	1.51 (0.34)	1.62 (0.38)	1.66 (0.47)
Diagnosis (81,84]	1.78 (0.30)	1.78 (0.37)	2.07 (0.37)	1.71 (0.36)	2.02 (0.33)	1.70 (0.37)	1.95 (0.46)
Diagnosis (84,87]	1.46 (0.31)	1.46 (0.38)	1.49 (0.39)	1.40 (0.38)	1.48 (0.34)	1.47 (0.38)	1.55 (0.48)
Diagnosis (87,90]	1.51 (0.31)	1.51 (0.38)	1.45 (0.39)	1.50 (0.37)	1.49 (0.36)	1.61 (0.38)	1.70 (0.47)
Diagnosis (90,93]	1.72 (0.30)	1.72 (0.38)	1.68 (0.38)	1.65 (0.37)	1.67 (0.34)	1.76 (0.37)	1.82 (0.47)
Diagnosis (93,96]	1.9 (0.30)	1.9 (0.37)	2.11 (0.37)	1.83 (0.36)	2.06 (0.33)	1.88 (0.37)	2.09 (0.46)
Diagnosis (96,99]	1.77 (0.30)	1.77 (0.37)	1.68 (0.38)	1.74 (0.37)	1.70 (0.35)	1.85 (0.37)	1.92 (0.46)
Diagnosis (99,102]	1.67 (0.30)	1.67 (0.38)	1.65 (0.38)	1.60 (0.37)	1.64 (0.34)	1.71 (0.38)	1.77 (0.47)
Diagnosis (102,105]	1.89 (0.30)	1.89 (0.37)	1.89 (0.38)	1.92 (0.36)	2.05 (0.34)	1.99 (0.37)	2.15 (0.46)
Diagnosis (105,108]	2.20 (0.29)	2.20 (0.36)	2.43 (0.36)	2.13 (0.36)	2.39 (0.32)	2.16 (0.37)	2.41 (0.45)
Diagnosis (108,111]	2.13 (0.29)	2.13 (0.37)	2.08 (0.37)	2.11 (0.36)	2.14 (0.34)	2.20 (0.37)	2.32 (0.46)
Diagnosis (111,114]	2.15 (0.29)	2.15 (0.36)	2.10 (0.37)	2.15 (0.36)	2.20 (0.34)	2.24 (0.37)	2.36 (0.46)
Diagnosis (114,117]	2.37 (0.29)	2.37 (0.36)	2.60 (0.36)	2.36 (0.36)	2.76 (0.32)	2.43 (0.36)	2.87 (0.45)
Diagnosis (117,120]	2.46 (0.29)	2.46 (0.36)	2.82 (0.36)	2.39 (0.36)	2.77 (0.31)	2.42 (0.36)	2.79 (0.45)
Diagnosis (120,123]	2.29 (0.22)	2.29 (0.36)	2.38 (0.37)	2.28 (0.36)	2.49 (0.33)	2.37 (0.36)	2.63 (0.45)
Diagnosis (123,126]	2.38 (0.29)	2.38 (0.36)	2.26 (0.37)	2.36 (0.36)	2.30 (0.33)	2.44 (0.36)	2.47 (0.45)
Diagnosis (126,129]	2.43 (0.29)	2.43 (0.36)	2.34 (0.37)	2.39 (0.36)	2.33 (0.32)	2.47 (0.36)	2.50 (0.45)
Diagnosis (129,132]	2.74 (0.29)	2.74 (0.36)	2.92 (0.36)	2.68 (0.35)	2.91 (0.32)	2.72 (0.36)	2.91 (0.44)
Diagnosis (132,135]	2.50 (0.29)	2.50 (0.36)	2.34 (0.37)	2.50 (0.36)	2.43 (0.33)	2.57 (0.36)	2.57 (0.45)
Diagnosis (135,138]	2.57 (0.29)	2.57 (0.36)	2.61 (0.36)	2.53 (0.35)	2.62 (0.32)	2.60 (0.36)	2.74 (0.45)
Diagnosis (138,141]	2.73 (0.29)	2.73 (0.36)	2.84 (0.36)	2.68 (0.35)	2.84 (0.32)	2.74 (0.36)	2.93 (0.44)
Diagnosis (141,144]	2.82 (0.29)	2.82 (0.36)	3.15 (0.35)	2.77 (0.35)	3.20 (0.31)	2.82 (0.36)	3.20 (0.44)
Diagnosis (144,147]	2.60 (0.29)	2.60 (0.36)	2.60 (0.36)	2.60 (0.35)	2.75 (0.32)	2.67 (0.36)	2.86 (0.45)
Diagnosis (147,150]	2.81 (0.29)	2.81 (0.36)	2.65 (0.36)	2.78 (0.35)	2.66 (0.32)	2.84 (0.36)	2.78 (0.44)
Diagnosis (150,153]	3.00 (0.28)	3.00 (0.35)	3.06 (0.36)	2.95 (0.35)	3.04 (0.31)	3.00 (0.36)	3.10 (0.44)
Diagnosis (153,156]	3.16 (0.28)	3.16 (0.35)	3.60 (0.35)	3.10 (0.35)	3.60 (0.31)	3.12 (0.36)	3.51 (0.44)
Diagnosis (156,159]	2.70 (0.29)	2.70 (0.36)	2.63 (0.36)	2.69 (0.35)	2.72 (0.32)	2.75 (0.36)	2.83 (0.45)
Diagnosis (159,162]	2.86 (0.28)	2.86 (0.36)	2.80 (0.36)	2.86 (0.35)	2.93 (0.32)	2.90 (0.36)	2.92 (0.44)
Diagnosis (162,165]	3.03 (0.28)	3.03 (0.35)	2.87 (0.36)	3.00 (0.35)	2.92 (0.31)	3.06 (0.36)	3.03 (0.44)
Diagnosis (165,168]	3.08 (0.28)	3.08 (0.35)	3.52 (0.35)	3.01 (0.35)	3.43 (0.31)	3.02 (0.36)	3.33 (0.44)
Diagnosis (168,171]	2.90 (0.28)	2.90 (0.35)	2.96 (0.36)	2.83 (0.35)	2.92 (0.31)	2.88 (0.36)	2.96 (0.44)
Diagnosis (171,174]	2.81 (0.29)	2.81 (0.36)	2.93 (0.36)	2.75 (0.35)	2.90 (0.32)	2.79 (0.36)	2.92 (0.44)
Diagnosis (174,177]	2.94 (0.28)	2.94 (0.35)	3.00 (0.36)	2.90 (0.35)	3.01 (0.32)	2.94 (0.36)	3.05 (0.44)
Diagnosis (177,180]	3.29 (0.28)	3.29 (0.35)	3.60 (0.35)	3.22 (0.35)	3.54 (0.30)	3.24 (0.36)	3.47 (0.44)
Diagnosis (180,183]	2.92 (0.28)	2.92 (0.35)	3.03 (0.36)	2.84 (0.35)	2.99 (0.31)	2.88 (0.36)	2.97 (0.44)
Diagnosis (183,186]	2.81 (0.28)	2.81 (0.36)	2.70 (0.36)	2.75 (0.35)	2.69 (0.31)	2.82 (0.36)	2.78 (0.44)
Diagnosis (186,189]	3.16 (0.28)	3.16 (0.35)	3.17 (0.35)	3.12 (0.35)	3.18 (0.31)	3.16 (0.36)	3.22 (0.44)
Diagnosis (189,192]	3.23 (0.28)	3.22 (0.35)	3.70 (0.35)	3.15 (0.35)	3.63 (0.30)	3.15 (0.36)	3.49 (0.44)
Diagnosis (192,195]	2.90 (0.28)	2.90 (0.35)	2.95 (0.36)	2.86 (0.35)	2.99 (0.31)	2.90 (0.36)	2.97 (0.44)
Diagnosis (195,198]	3.05 (0.28)	3.05 (0.35)	2.99 (0.36)	2.98 (0.35)	2.97 (0.31)	3.02 (0.36)	2.94 (0.44)
Diagnosis (198,201]	2.94 (0.28)	2.94 (0.35)	3.01 (0.36)	2.87 (0.35)	2.98 (0.31)	2.91 (0.36)	2.98 (0.44)
Diagnosis (201,204]	3.26 (0.28)	3.26 (0.35)	3.61 (0.35)	3.19 (0.35)	3.55 (0.31)	3.19 (0.36)	3.41 (0.44)
Diagnosis (204,207]	2.92 (0.28)	2.92 (0.35)	3.00 (0.36)	2.87 (0.35)	2.96 (0.31)	2.89 (0.36)	2.94 (0.44)
Diagnosis (207,210]	2.84 (0.29)	2.84 (0.36)	2.76 (0.36)	2.80 (0.35)	2.77 (0.32)	2.85 (0.36)	2.83 (0.44)
Diagnosis (210,213]	3.11 (0.28)	3.11 (0.35)	2.94 (0.36)	3.06 (0.35)	2.94 (0.31)	3.11 (0.36)	3.0 (0.44)
Diagnosis (213,216]	3.08 (0.28)	3.08 (0.35)	3.46 (0.35)	3.01 (0.35)	3.38 (0.31)	3.00 (0.36)	3.20 (0.44)
Diagnosis (216,219]	2.95 (0.28)	2.95 (0.35)	2.89 (0.36)	2.88 (0.35)	2.88 (0.31)	2.92 (0.36)	2.88 (0.44)
Diagnosis (219,222]	2.84 (0.29)	2.84 (0.36)	2.52 (0.36)	2.84 (0.35)	2.63 (0.32)	2.90 (0.36)	2.72 (0.44)
Diagnosis (222,225]	3.04 (0.28)	3.04 (0.35)	2.90 (0.36)	3.02 (0.35)	2.93 (0.31)	3.07 (0.36)	3.018 (0.44)
Diagnosis (225,228]	3.19 (0.28)	3.19 (0.35)	3.46 (0.35)	3.12 (0.35)	3.40 (0.31)	3.13 (0.36)	3.31 (0.44)
Diagnosis (228,231]	2.89 (0.29)	2.89 (0.35)	2.95 (0.36)	2.82 (0.35)	2.26 (0.31)	2.87 (0.36)	2.95 (0.44)
Diagnosis (231,234]	2.92 (0.28)	2.92 (0.35)	2.89 (0.36)	2.91 (0.35)	3.02 (0.32)	2.96 (0.36)	3.11 (0.44)
Diagnosis (234,237]	2.91 (0.28)	2.91 (0.35)	2.94 (0.36)	2.87 (0.35)	2.99 (0.31)	2.93 (0.36)	3.05 (0.44)
Diagnosis (237,240]	3.11 (0.28)	3.11 (0.35)	3.37 (0.35)	3.04 (0.35)	3.32 (0.31)	3.08 (0.36)	3.34 (0.44)
Diagnosis (240,243]	2.88 (0.28)	2.88 (0.35)	2.82 (0.36)	2.84 (0.35)	2.83 (0.32)	2.89 (0.36)	2.92 (0.44)
Diagnosis (243,246]	2.80 (0.29)	2.80 (0.36)	2.89 (0.36)	2.73 (0.35)	2.87 (0.31)	2.77 (0.36)	2.89 (0.44)
Diagnosis (246,249]	2.88 (0.29)	2.88 (0.35)	2.91 (0.36)	2.82 (0.35)	2.89 (0.31)	2.88 (0.36)	2.97 (0.44)
Diagnosis (249,252]	3.02 (0.28)	3.02 (0.35)	3.30 (0.35)	2.95 (0.35)	3.24 (0.31)	2.98 (0.36)	3.24 (0.44)
Diagnosis (252,255]	2.76 (0.29)	2.76 (0.36)	2.71 (0.36)	2.71 (0.35)	2.70 (0.31)	2.77 (0.36)	2.79 (0.44)
Diagnosis (255,258]	2.52 (0.29)	2.52 (0.36)	2.48 (0.37)	2.51 (0.36)	2.58 (0.32)	2.55 (0.36)	2.60 (0.45)
Diagnosis (258,261]	2.79 (0.29)	2.80 (0.36)	2.74 (0.36)	2.75 (0.35)	2.74 (0.32)	2.1 (0.36)	2.83 (0.44)
Diagnosis (261,264]	3.08 (0.28)	3.08 (0.35)	3.30 (0.35)	3.02 (0.35)	3.25 (0.31)	3.03 (0.36)	3.17 (0.44)
Diagnosis (264,267]	2.66 (0.29)	2.66 (0.36)	2.58 (0.36)	2.60 (0.35)	2.57 (0.32)	2.66 (0.36)	2.63 (0.45)
Diagnosis (267,270]	2.57 (0.29)	2.57 (0.36)	2.50 (0.37)	2.51 (0.35)	2.48 (0.32)	2.55 (0.36)	2.47 (0.45)
Diagnosis (270,273]	2.61 (0.29)	2.61 (0.36)	2.56 (0.37)	2.58 (0.35)	2.62 (0.33)	2.64 (0.36)	2.72 (0.45)
Diagnosis (273,276]	2.90 (0.29)	2.90 (0.36)	3.09 (0.36)	2.83 (0.35)	3.05 (0.31)	2.87(0.36)	3.08 (0.44)
Diagnosis (276,279]	2.48 (0.29)	2.48 (0.36)	2.32 (0.37)	2.50 (0.35)	2.47 (0.33)	2.56 (0.36)	2.57 (0.45)

(Table continues...)

Table 5.3: Count regression models using as predictors the delays quarter and diagnosis quarters

	Poisson	Poisson Overdispersion	Negative Binomial	Zero-Inflated Poisson	Zero-Inflated Negative Binomial	Hurdle Poisson	Hurdle Negative Binomial
	value (sd)	value (sd)	value (sd)	value (sd)	value (sd)	value (sd)	value (sd)
Diagnosis (279,282]	2.63 (0.29)	2.63 (0.36)	2.62 (0.37)	2.59 (0.35)	2.64 (0.33)	2.65 (0.36)	2.73 (0.45)
Diagnosis (282,285]	2.54 (0.29)	2.54 (0.36)	2.54 (0.37)	2.48 (0.35)	2.53 (0.32)	2.52 (0.36)	2.55 (0.45)
Diagnosis (285,288]	2.77 (0.29)	2.76 (0.36)	2.83 (0.36)	2.70 (0.35)	2.81 (0.32)	2.72 (0.36)	2.77 (0.44)
Diagnosis (288,291]	2.51 (0.29)	2.51 (0.36)	2.33 (0.37)	2.52 (0.36)	2.49 (0.33)	2.58 (0.36)	2.58 (0.45)
Diagnosis (291,294]	2.27 (0.29)	2.27 (0.36)	2.03 (0.38)	2.22 (0.36)	2.05 (0.33)	2.29 (0.37)	2.15 (0.45)
Diagnosis (294,297]	2.51 (0.29)	2.51 (0.36)	2.38 (0.37)	2.47 (0.36)	2.39 (0.33)	2.52 (0.36)	2.48 (0.45)
Diagnosis (297,300]	2.76 (0.29)	2.76 (0.36)	2.75 (0.37)	2.69 (0.35)	2.73 (0.32)	2.73 (0.36)	2.72 (0.45)
Diagnosis (300,303]	2.48 (0.29)	2.48 (0.36)	2.23 (0.38)	2.45 (0.36)	2.25 (0.33)	2.50 (0.36)	2.36 (0.45)
Diagnosis (303,306]	2.27 (0.29)	2.27 (0.36)	2.18 (0.38)	2.26 (0.36)	2.29 (0.34)	2.32 (0.37)	2.38 (0.46)
Diagnosis (306,309]	2.30 (0.29)	2.30 (0.36)	2.26 (0.38)	2.24 (0.36)	2.25 (0.33)	2.28 (0.37)	2.28 (0.45)
Diagnosis (309,312]	2.41 (0.29)	2.41 (0.36)	2.30 (0.38)	2.36 (0.36)	2.31 (0.33)	2.42 (0.36)	2.41 (0.45)
Diagnosis (312,315]	2.34 (0.29)	2.34 (0.36)	2.12 (0.39)	2.30 (0.36)	2.15 (0.33)	2.36 (0.36)	2.27 (0.46)
Diagnosis (315,318]	2.22 (0.29)	2.22 (0.37)	1.95 (0.39)	2.24 (0.36)	2.13 (0.35)	2.29 (0.37)	2.23 (0.46)
Diagnosis (318,321]	2.25 (0.29)	2.25 (0.37)	2.14 (0.39)	2.24 (0.36)	2.25 (0.35)	2.29 (0.37)	2.34 (0.46)
Diagnosis (321,324]	2.62 (0.29)	2.61 (0.36)	2.46 (0.39)	2.56 (0.36)	2.48 (0.34)	2.61 (0.36)	2.54 (0.45)
Diagnosis (324,327]	2.46 (0.29)	2.46 (0.36)	2.27 (0.40)	2.41 (0.36)	2.30 (0.34)	2.46 (0.36)	2.36 (0.46)
Diagnosis (327,330]	2.17 (0.30)	2.17 (0.37)	1.82 (0.41)	2.12 (0.36)	1.88 (0.36)	2.17 (0.37)	1.95 (0.47)
Diagnosis (330,333]	2.21 (0.30)	2.21 (0.37)	198 (0.42)	2.16 (0.36)	2.02 (0.36)	2.21 (0.37)	2.09 (0.47)
Diagnosis (333,336]	2.25 (0.30)	2.251 (0.37)	2.03 (0.44)	2.20 (0.36)	2.07 (0.38)	2.25 (0.37)	2.14 (0.48)
Diagnosis (336,339]	2.04 (0.31)	2.04 (0.38)	2.01 (0.5)	2.00 (0.37)	2.04 (0.44)	2.04 (0.38)	2.09 (0.52)
R^2 or pseudo R^2	0.99	0.97	0.88	0.95	0.89	0.91	0.85
RMSD	3.29	3.29	7.64	4.16	6.61	7.36	7.85
AIC	9768.9	9768.9	9433.5	9819.82	9578.59	10609.55	10432.34
BIC	10603.82		10274.51	11265.19	11029.24	12055.19	11883.24
Number parameters	138	138	139	239	240	239	240
Log L	-4746		-4578	-4671	-4549	-5066	-4976

From all measures of goodness-of-fit of the models, the Poisson process represented the best choice.

Harris described the basic statistical model in [121], and an application to the Portuguese AIDS incidence data reported until 2001 was presented in [203, 204]. This process may be formulated as a bivariate Poisson process corresponding to a log-linear model with an independence structure, and can be used as a measure of the stationarity of the process.

Assume that X_{ij} are independent and identically distributed following a Poisson distribution with parameter μ_{ij}. If $Y^* = \sum_{i=1}^{m} X_{ij}$ represents the total number of AIDS cases diagnosed in $[t_{j-1}, t_j]$, whether or not they have been reported by the time t_n, then $\{Y^*\}$ are independent and identically distributed, following a Poisson distribution with mean $\sum_{i=1}^{m} \mu_{ij}(\theta)$, which may be interpreted as the AIDS incidence in the time interval $[t_{j-1}, t_j]$.

For each $j = 1, \ldots, n$, we also have that $\{X_{ij}, X_{2j}, \ldots, X_{mj}\} | Y_j^*$ has a multinomial distribution upon Y_j^* trials with probabilities

$$\Pi = \{\Pi_{1j}(\theta), \Pi_{2j}(\theta), \ldots, \Pi_{mj}(\theta)\}. \tag{5.2}$$

Here

$$\Pi_{ij}(\theta) = \frac{\mu_{ij}}{\sum_{i=1}^{m} \mu_{ij}(\theta)} \tag{5.3}$$

is the probability of an AIDS case being reported with a delay in the time interval $[d_{i-1}, d_i]$, given that it has been diagnosed in the time interval $[t_{j-1}, t_j]$.

Assuming a separability restriction on $\mu_{ij}(\theta)$, the parameter θ can be partitioned as (α, θ) so that

$$\mu_{ij}(\alpha, \beta) = \Pi_{ij}(\alpha) \sum_{i=1}^{m} \mu_{ij}(\beta). \tag{5.4}$$

As such, the delay probabilities depend only on α and the incidence depends only on β.

For each $j = 1, \ldots, n$, let

$$Y_j = \sum_{i=1}^{m_j} X_{ij} \tag{5.5}$$

be the number of observed AIDS cases that were diagnosed at the time interval $[T_{j-1}, T_j]$ and reported until the end of the observation period. The $\left\{Y_j | Y_j^*\right\}$ are independent and identically distributed with a Binomial distribution with the probability of a success given by $\Omega_j(\alpha) = \sum_{i=1}^{m_j} \Pi_{ij}(\alpha)$.

This represents the probability of an AIDS case being reported until T_n, given that it was diagnosed at the time interval $[T_{j-1}, T_j]$.

For $j = 1, \ldots, n$, $\{(X_{1j}, X_{2j}, \ldots, X_{mj}) | Y_j\}$ is multinomial upon Y_j trials with probabilities

$$\left\{ \frac{\Pi_{1j}(\alpha)}{\Omega_j(\alpha)}, \ldots, \frac{\Pi_{m_j j}(\alpha)}{\Omega_j(\alpha)} \right\}. \tag{5.6}$$

Here, $\frac{\Pi_{ij}(\alpha)}{\Omega_j(\alpha)}$ is the probability that an AIDS case is reported with delay falling in $[d_{i-1}, d_i]$, given that it is diagnosed in $[T_{j-1}, T_j]$ and reported until T_n.

The likelihood function $L(\theta)$, given the separability restriction above, can be written as

$$L(\alpha, \beta) = L_c(\alpha) L_m(\beta) \tag{5.7}$$

with $L_c(\alpha)$ being the conditional likelihood based on the observed data and $L_m(\beta)$ the marginal likelihood. As such,

$$L_c(\alpha) = \prod_{j=1}^{n} \prod_{i=1}^{m_j} \left[\frac{\pi_{ij}(\alpha)}{\Omega_j(\alpha)} \right]^{x_{ij}} \tag{5.8}$$

and

$$L_m(\alpha, \beta) = \prod_{j=1}^{n} \left[\Omega_j(\alpha) \sum_{i=1}^{m} \mu_{ij}(\beta) \right]^{y_j} \exp\left(-\Omega_j(\alpha) \sum_{i=1}^{m} \mu_{ij}(\beta) \right). \tag{5.9}$$

Brookmeyer and Damiano relied in the likelihood (5.8) to produce the estimates for the parameters in the reporting delay distribution [203, 204, 212, 199, 208]. They considered that $\left\{X_{1j}, X_{2j}, \ldots, X_{m_j j}\right\} | Y_j$ follows a Multinomial distribution with probabilities given by

$$p_{ij}(\alpha_i) = \frac{\Pi_{ij}(\alpha_i)}{\Omega_j(\alpha_i)} = \frac{\exp(\alpha_i)}{\sum_{i=1}^{m_j} exp(\alpha_i)} \tag{5.10}$$

which can be easily estimated using a Multinomial-Poisson transformation and standard regression packages. As mentioned by Amaral, Pereira and Paixão [203, 204], the number of diagnosed and reported cases in the time calendar $[T_{j-1}, T_j]$, $j = 1, \ldots, n$ can be multiplied by the weighting factor

$$\left[\sum_{i=1}^{m_j - 1} p_{ij}(\hat{\alpha_i}) + \frac{p_{m_j j}\left(\hat{\alpha_{m_j}}\right)}{2} \right]^{-1} \tag{5.11}$$

in order to obtain an estimate of the cases which were diagnosed and reported [204, 208].

A common consideration is letting $\alpha_1 = \beta_1 = 0$, which is equivalent to say that at the beginning of the epidemic there was only one person [208].

Returning to Table 5.3, a further inspection of the fitted Poisson model was performed. The relationship between the Poisson mean and variance presents slight problems, represented on Figure 5.31, and few outliers seem to be affecting the fit of the model - Figure

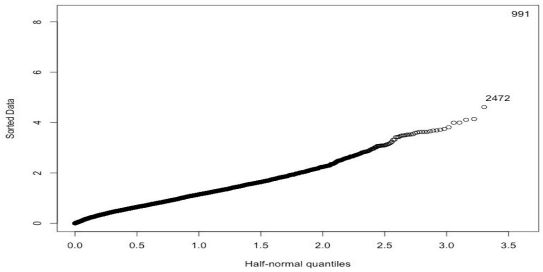


Figure 5.30: Half-normal plot of the residuals of the Poisson model

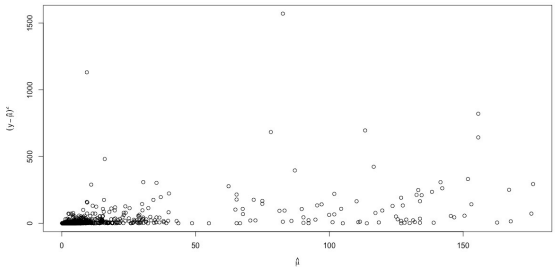


Figure 5.31: Relationship between mean and variance

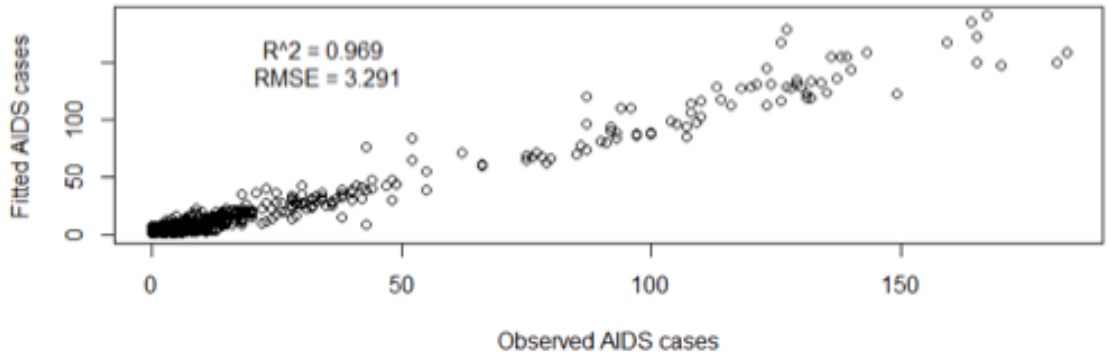


Figure 5.32: Number of observed AIDS cases against the predicted number of AIDS by the Poisson model

5.30. The observed problems of the relationship between the mean and variance of the Poisson process seem to be related to the existence of missing cells in Table 5.2 [176]. For this model, it was obtained a deviance of 4276.5 with 2997 degrees of freedom, indicating a non-stationary process or quasi-independent model which will be further analysed in subsection 5.2.2.

However, the comparison of the observed values against the predicted values suggests the model is useful in capturing the observed patterns. For further evaluating the accuracy of the model, the holdout method was applied. The X_{ij} data was randomly divided into two sets: 75% of the data was used to train the model (2351 cases) and 25% used to test it (784 cases). Within the training set, the maximum value of the log-likelihood function for the estimations was LL = -4746. Within the test set, it was estimated $RMSE = 5.511$ and $R^2 = 0.899$. These results indicate that the Poisson model is able to capture the trend in the data and fits moderately well. For obtaining the final reporting delay estimate we used the whole data set; $RMSE = 3.291$ and $R^2 = 0.969$ were found.

The results from estimating the reporting delay probabilities with equation (5.10) are presented in Table 5.4 .

These estimates are close to the ones obtained previously for the Portuguese epidemic in [203, 204].

Table 5.4: Estimated reporting delay probabilities for each quarter

Reporting delay in quarters	Estimated probabilities of reporting delay
0 - 3	0.548
4 - 6	0.125
7 - 9	0.062
10 - 12	0.039
13 - 15	0.030
16 - 18	0.021
19 - 21	0.018
22 - 24	0.014
25 - 27	0.013
28 - 30	0.011
31 - 33	0.010
34 - 36	0.010

The evaluation indicates that approximately 80% of the AIDS cases were reported within one year after the diagnosis and that the majority of the cases were notified within the first three months. The accuracy of the procedure can also be indicated by the relative stability over time of the resulting estimates of the reporting delay observed in Figure 5.21.

Underreporting

Fader and Hardie in 2000 proposed a model for under-reported Poisson counts using natural conjugate prior distributions and maximum likelihood, described in [213]. The methodology proposed that the probability of an AIDS case not being reported can be estimated by the Beta Binomial / Negative Binomial model developed by [213].

Let Y^* be the current (unobserved) number of events and Y be the number of reported events. We are interested in the following:

- the probability of occurring y^* AIDS cases, conditional on the fact that y AIDS cases have been reported:
$$P\left(Y^* = y^* | Y = y\right); \tag{5.12}$$
- the distribution of the reporting probability of one case in light of the fact that y AIDS cases have been reported:
$$g\left(p | Y = y\right); \tag{5.13}$$
- the distribution of the true rate parameter, conditional on y reported AIDS cases:
$$f\left(\lambda | Y = y\right). \tag{5.14}$$

The model assumes that:

(i) The actual (unobserved) number of AIDS cases - Y^* - in the unit time interval is Poisson distributed with rate λ, i.e.,

$$Y^* | \lambda \sim P\left(\lambda\right) \tag{5.15}$$

$$P\left(Y^* = y^* | \lambda\right) = \frac{\lambda^{y^*} \exp(-\lambda)}{y^*!}, y^* = 0, 1, \ldots; \lambda > 0 \tag{5.16}$$

(ii) An AIDS case is reported with probability p. So, conditional on y^*, the number of reported AIDS, Y, follows a binomial distribution with probability mass function

$$Y|Y^* = y^* \sim B\left(y^*, p\right) \tag{5.17}$$

$$P\left(Y = y|y^*, p\right) = \binom{y^*}{y} p^y \left(1-p\right)^{y^*-y}, y = 0, 1, \ldots, y^*; 0 < p < 1, y^* \in \mathbb{Z}^+ \tag{5.18}$$

(iii) The probability of an AIDS case being reported (p) is independent from the rate development of AIDS (λ)

(iv) $\lambda \sim Gamma\left(\alpha, r\right)$ (the $Gamma$ function is a conjugate prior of the Poisson family)

$$f\left(\lambda\right) = \frac{\alpha^r \lambda^{r-1} \exp(-\lambda\alpha)}{\Gamma\left(r\right)}, \lambda > 0; r, \alpha > 0 \tag{5.19}$$

$$\text{and } \Gamma\left(r\right) = \int_0^\infty t^{r-1} \exp\left(t\right) dt \tag{5.20}$$

(v) $p \sim Beta\left(a, b\right)$

$$g\left(p\right) = \frac{1}{B(a,b)} p^{a-1} \left(1-p\right)^{b-1}, 0 < p < 1, a, b > 0 \tag{5.21}$$

$$\frac{a}{B\left(a,b\right)} = \frac{\Gamma\left(a+b\right)}{\Gamma\left(a\right)\Gamma\left(b\right)} \tag{5.22}$$

From assumptions (i) and (iii) it follows that

$$P\left(Y^* = y^*\right) = \frac{\Gamma\left(r+y^*\right)}{\Gamma\left(r\right) y^*!} \left(\frac{\alpha}{\alpha+1}\right) \left(\frac{1}{\alpha+1}\right)^{y^*} \tag{5.23}$$

and assumptions (ii) and (iv) define the Beta-Binomial (BB) model

$$Y|Y^* \sim BB(\alpha, \beta) \tag{5.24}$$

with

$$\begin{aligned} P\left(Y = y|Y^*\right) &= \binom{y^*}{y} \frac{Beta\left(\alpha+y, \beta+y^*-y\right)}{Beta\left(\alpha, \beta\right)} \\ &= \binom{y^*}{y} \frac{\Gamma\left(\alpha+y\right)\Gamma\left(\beta+y^*-y\right)}{\Gamma\left(\alpha+\beta+y^*\right)} \frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}. \end{aligned} \tag{5.25}$$

From assumptions (i) and (ii), $Y|\lambda, p \sim P\left(\lambda p\right)$ i.e.,

$$\begin{aligned} P\left(Y = y|\lambda, p\right) &= \sum_{Y^*=y}^{\infty} P\left(Y = y|n, p\right) P\left(Y^* = y^*|\lambda\right) \\ &= \frac{\left(\lambda p\right)^y \exp\left(-\lambda p\right)}{y!}. \end{aligned} \tag{5.26}$$

Finally, from (i), (iii), (iv) and (v):

$$\begin{aligned} P(Y=y) &= \int\int P(Y=y|\lambda,p)\, f(\lambda)\, g(p)\, d\lambda dp \\ &= \frac{\Gamma(r+y)}{\Gamma(r)\, y!}\left(\frac{\alpha}{\alpha+1}\right)^r \left(\frac{1}{\alpha+1}\right)^y \frac{\Gamma(a+y)}{\Gamma(a)} \frac{\Gamma(a+b)}{\Gamma(a+b+y)} \times \\ &\quad \times {}_2F_1\left(r+y, b; a+b+y; \frac{1}{\alpha+1}\right) \\ &= \frac{1}{(r+y)\, Beta(r, y+1)} \frac{Beta(a+y; b)}{Beta(a,b)} \frac{\alpha^r}{(\alpha+1)^{r+y}} \times \\ &\quad \times {}_2F_1\left(r+y, b; a+b+y; \frac{1}{\alpha+1}\right) \\ &\quad , y = 0, 1, 2, \ldots; r, \alpha, a, b > 0 \end{aligned} \tag{5.27}$$

where ${}_2F_1(.)$ is the Gauss hypergeometric function defined by ${}_2F_1(a, b; c; z) = \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{(c)_n\, n!} z^n$, with $|z| < 1, c \neq 0, -1, -2, \ldots$. The symbol $(a)_n$ is the Pochhammer's symbol and can be calculated using

$$(a)_n = \frac{\Gamma(a+n)}{\Gamma(a)} \tag{5.28}$$

This is known as the BB / NBD model with

$$\begin{aligned} E(Y) &= \frac{ra}{\alpha(a+b)} \\ Var(Y) &= \frac{ra}{\alpha(a+b)}\left[\frac{1}{\alpha(a+b+1)}\left(\frac{rb}{a+b} + a + 1\right) + 1\right]. \end{aligned}$$

It can be shown ([213, 214, 215, 216]) that the expected value of Y^* given Y is

$$E(Y^*|Y=y) = y + \frac{r+y}{\alpha+1} \frac{Beta(a+y, b+1)}{Beta(a+y, b)} \frac{{}_2F_1\left(r+y+1, b+1; a+b+y+1; \frac{1}{\alpha+1}\right)}{{}_2F_1\left(r+y, b; a+b+y; \frac{1}{\alpha+1}\right)}, \tag{5.29}$$

that the expected value of π conditional to Y is

$$E(\pi|Y=y) = \frac{a+y}{a+b+y} \frac{{}_2F_1\left(r+y, b; a+b+y+1; \frac{1}{\alpha+1}\right)}{{}_2F_1\left(r+y, b; a+b+y; \frac{1}{\alpha+1}\right)} \tag{5.30}$$

and finally, that the expected value of λ given Y is

$$E(\lambda|Y=y) = \frac{r+y}{\alpha+1} \frac{{}_2F_1\left(r+y+1, b; a+b+y; \frac{1}{\alpha+1}\right)}{{}_2F_1\left(r+y, b; a+b+y; \frac{1}{\alpha+1}\right)} \tag{5.31}$$

The estimation of a, b, r and α from the BB/NBD model can be obtained by minimizing the log likelihood function which, in this case, is a complex numerical problem since this function has several ridges.

To study which combination of initial parameters and which numerical optimization algorithms are able to find the minimum, the function was reparametrized in order to expand the search space. A combination of initial parameters in the set $(r, \alpha, a, b) \in [0,1] \times [0,1] \times [0,1] \times [0,1]$ with $step = 0.1$ was considered; that is, a web of $10 \times 10 \times 10 \times 10 = 10000$ starting points was created. The minimum was found by application of the most common optimization algorithms, namely:

- the Nelder and Mead algorithm used for unconstrained problems (no lower or upper specified); it uses only function values, is robust but relatively slow and works reasonably well for non-differentiable functions (Figure 5.33a);
- the BFGS or the variable metric method; updates an approximation of the inverse Hessian using the BFGS update formulas, along with an acceptable point line search strategy but appears to work best with analytic gradients (Figure 5.33b);
- the L-BFGS method; it is a generalization of the BFGS allowing box constraints, ie, each variable can be given a lower and/or upper bound (Figure 5.33c);
- the BOBYQA algorithm; it implements an optimization by quadratic approximation for box constrained problems (Figure 5.33d);
- the NEWUOA algorithm; it implements an optimization by quadratic approximation for unconstrained minimization (Figure 5.33e);
- the NLM algorithm; it implements a Newton-type algorithm (Figure 5.33f);
- the NLMINB algorithm; it implements unconstrained and box-constrained optimization using PORT routines (Figure 5.33g);
- the SPG algorithm; it implements a spectral projected gradient method for large-scale optimization with simple constraints (Figure 5.33h);
- the UCMINF algorithm; used for general-purpose unconstrained non-linear optimization (Figure 5.33i).

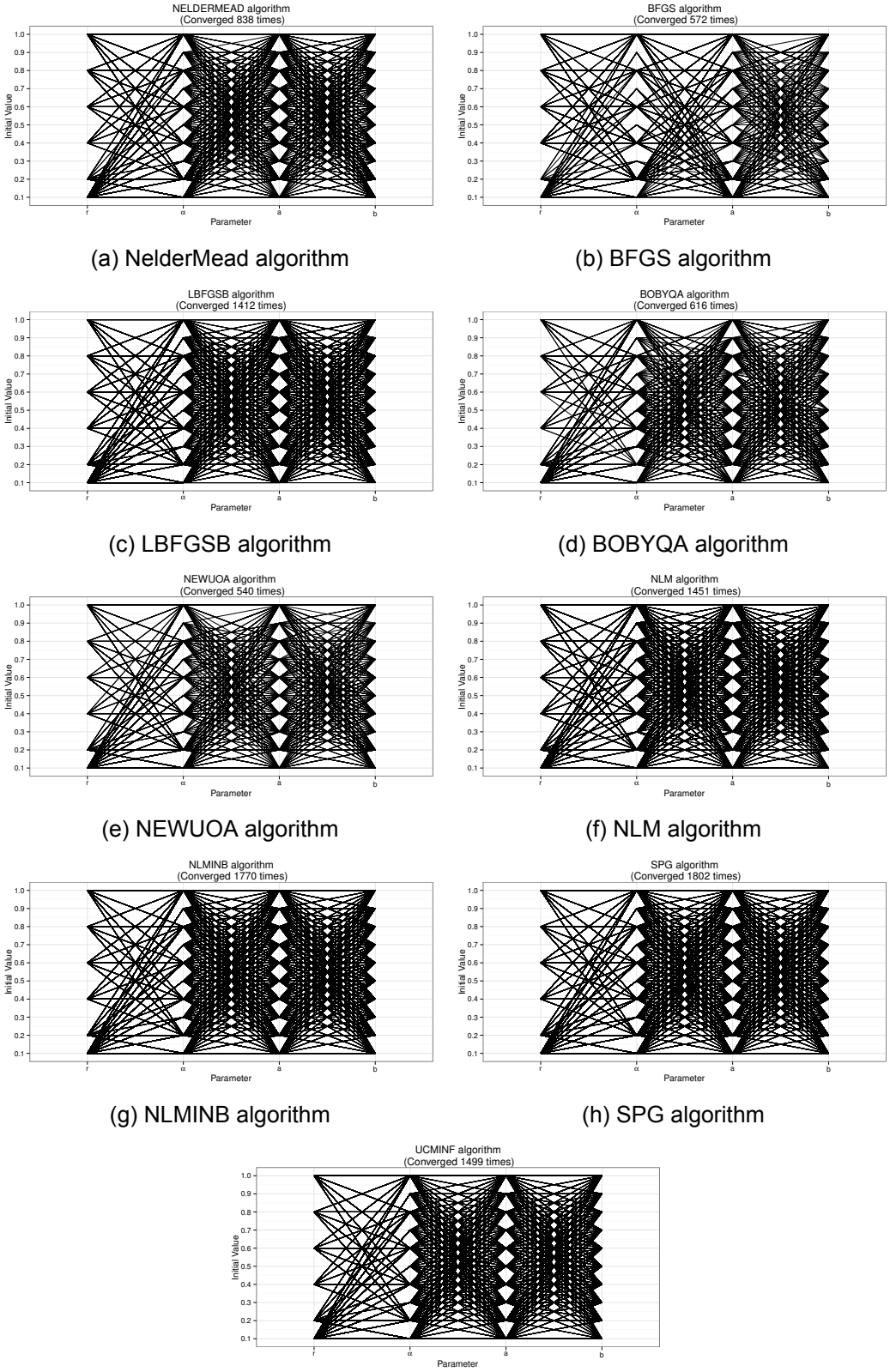
From the 10000 times the optimization algorithms were started, the algorithm SPG converged more times, that is 1770 (almost 18% of the runs) and the NEWUOA converged the least times, 540 (5.4%). The algorithms NLMINB, UCMINF and LBFGSB performed close to the SPG.

In terms of starting values naturally, the Neldermead, NLMINB, SPG, UCMINF, NLM and LBFGSB algorithms allowed a larger variety of combinations but a pattern of combinations is common to all algorithms; that is, r must be in $\{1; 0.8; 0.6; 0.4; 0.2\}$ and α must be in the same subset (here, r and α are the Γ parameters). From the different methods, a maximum value for the log-likelihood function of $LL = -586.601$ was found, corresponding to $\hat{r} = 61.86$, $\hat{\alpha} = 0.21$, $\hat{a} = 1.20$ and $\hat{b} = 0.87$. To evaluate the goodness-of-fit of the model, the number of AIDS cases was divided into intervals of length 25 and the χ^2- test was used. The values $\chi^2 = 9.03$ and $p = 0.43$ were obtained, thus supporting the use of the model (5.34). Using equation (5.30), the model suggests that the probability of an AIDS case being notified is higher when a large amount of AIDS cases had already been notified. This relationship is represented in Figure 5.35. The estimated number of diagnosed AIDS cases adjusted for reporting delay and underreporting was estimated by equation (5.29). As Figure 5.36 suggests, this distribution is very different from the distribution of AIDS cases diagnosed and reported to CVDT, not capturing the marked decrease of the AIDS incidence from 1999 onwards. Although the model goodness-of-fit seems to be reasonable, the underreporting process is not being captured using only marginal data.

5.2.2 Separate Modelling - Accounting for a Non Stationary Process

As the need for medical care is much higher in late-stage infections, the under - diagnosis is less likely to occur in AIDS cases. We will therefore restrict ourselves only to this stage in the following analysis.

The reporting delay is defined as the time mediating from HIV - AIDS diagnosis to the reporting of this event at national level [69].



(a) NelderMead algorithm

(b) BFGS algorithm

(c) LBFGSB algorithm

(d) BOBYQA algorithm

(e) NEWUOA algorithm

(f) NLM algorithm

(g) NLMINB algorithm

(h) SPG algorithm

(i) UCMINF algorithm

Figure 5.33: Optimization Performance

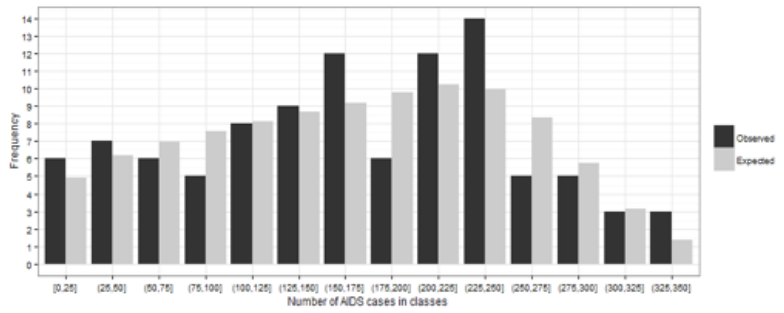


Figure 5.34: Number of AIDS cases observed against number of AIDS cases estimated by the BB/NBD model.

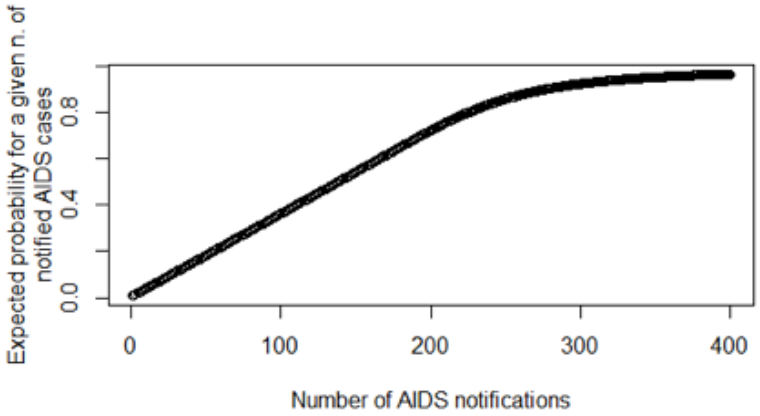


Figure 5.35: Expected probability of an AIDS case being notified for a given number of notified AIDS cases. The estimations were obtained with the BB/NBD model

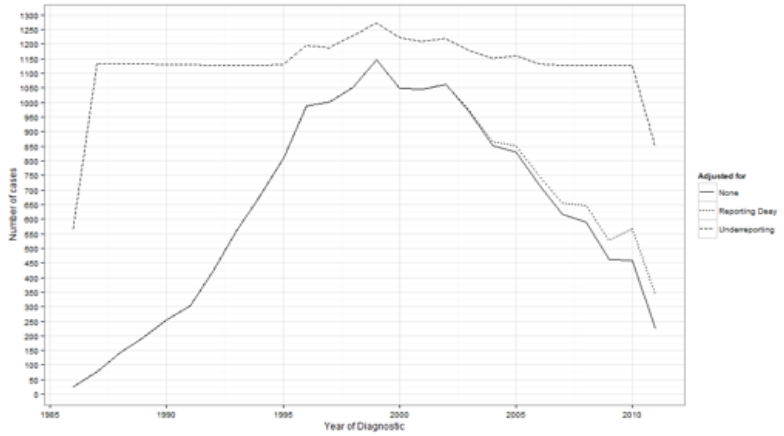


Figure 5.36: Number of AIDS cases diagnosed and reported to CVDET per year (without adjustment) and number of AIDS cases adjusted for reporting delay and underreporting

The first case diagnosed (and reported) will be represented by 0 and x^* will denote the end of the observation period. The time interval $[0, x^*]$ thus represents the observed time for diagnosis observed.

Let this time interval be divided into 3 months-unit length. The same division is set to the reporting delay time interval.

The AIDS cases are then cross-classified by the diagnosis and reporting delay quarter.

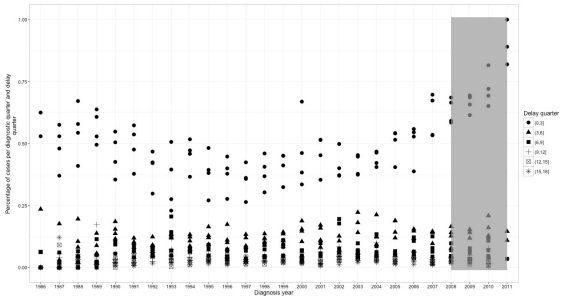


Figure 5.37: Percentage of AIDS cases per diagnosis and delay quarters. The shaded region identifies those (recent) years that have to be corrected.

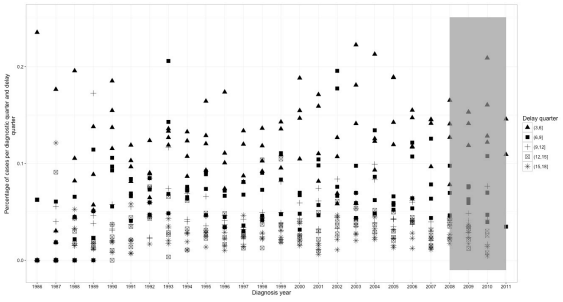


Figure 5.38: Percentage of AIDS cases per quarter of diagnosis and delay quarter longer than 3 months

We denote by X_{ij} the annual percentage of AIDS cases that were diagnosed in quarter i and have a reporting delay falling into quarter j. Fixing a reporting delay quarter, one can observe a delay pattern over time.

We now proceed as follows:

1. identification of homogeneous patterns for the observed delays considering that there is no bias. The clustering is performed by a longitudinal version of the K-means algorithm. This step identifies the tendencies across the reporting delays, thus reducing the number of curves of reporting delays to be considered.
2. evaluation of the effects of the reporting delays and date of diagnoses on the evolution of the epidemic (percent-wise). We use linear regression with estimation performed by generalized least squares, thus allowing for heteroscedastic errors. We will consider only cases diagnosed until 2008 due to the reporting delay.
3. Imputation on X_{ij} according to the obtained results.

Due to the lack of understanding of the HIV infection development during its first years, we will consider only cases diagnosed after June 1986.

Longitudinal K-means Figure 5.37 depicts the reporting delays that are registered in the national AIDS surveillance system. The delays were grouped into trimesters and the annual percentage of cases within each diagnosis year is represented.

The most recent year (shaded region) does not seem to be describing the real situation as several cases have not been notified yet. Most of the cases are reported within 3 months after diagnosis but some are still being reported with more than one year of delay. For the sake of clarity, delays longer than 18 months are omitted from figures 5.37 and 5.38. The longer delays, e.g. more than 6 months, have an almost constant behaviour through time not exceeding 10% of the cases (figure 5.38). Graphical inspection may suggest a 2-cluster structure since the delay curve $(0, 3]$ is fully separated from the rest. However, due to epidemiological interest it is important to explore other clustering structures.

Since clustering analysis involves exploratory methods, 2-, 3- and 4-cluster structures, each of them with 40 randomly chosen starting points, were studied. The longitudinal K-means algorithm was applied, considering the euclidean distance with the Gower adjustment and implemented via Expectation-Maximization. The optimal number of clusters was determined with the help of several criteria: Calinsky and Harabatz, Ray and Turi, Davies and Bouldin, and the Bayesian and Akaike Information Criteria (BIC and AIC, respectively).

All should ideally agree and be as large as possible except BIC and AIC that should be low [182].

For the reporting delays, and considering partitions from 2 to 4 clusters, the behaviour of the quality criteria is represented in figure 5.39. The clustering result for 3 classes is described in table 5.5.

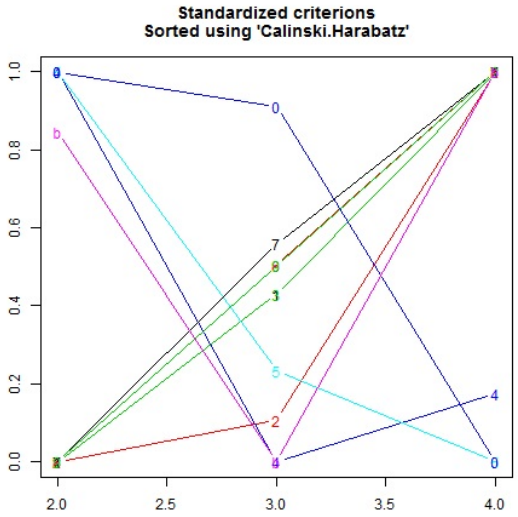


Figure 5.39: Quality Criteria for longitudinal clustering. 0 - Calinsky and Harabatz; 1 - Calinsky and Harabatz2; 3 - Calinsky and Harabatz3; 4 -Ray and Turi; 5 -Davies and Bouldin; 6 - Bayesian Information Criterion (BIC); 7 - BIC with correction for finite sample size; 8- Akaike Information Criterion (AIC); 9- AIC with correction for finite sample size

Table 5.5: Description of analysed clusters

Number of Clusters	Clusters ID
2	C1: (0,3]; C2: $>$3 months
3	C1: (0,3]; C2: (3,6]; C3: $>$6 months
4	C1: (0,3]; C2: (3,6]; C3: (6,9]; C4: $>$9 months

From the analysis of figure 5.39 and table 5.5, we considered a 2-cluster structure.

Generalised least squares Since the number of AIDS cases in the most recent diagnosis are biased due to the effect of the reporting delay, the latest 4 years were removed from the analysis.

The general equation of the adjusted time evolution model for the reporting delay curves is

$$\widehat{X}_i \quad = \quad 0.461 + 0.0104i + (-0.415)C2 + (-0.0103)i * C2 \tag{5.32}$$

were X_i is the annual percentage of AIDS cases, i is the year of diagnosis and $C2$ is as described in Table 5.5 (the reference level is the cluster $(0, 3]$) .

The adjusted time effect on the AIDS incidence rate has to be described within each of the identified clusters: in Cluster $[0, 3]$ months (resp. $[3 - 18]$ months), the model predicted a yearly increase of 1.04% (p<0.001) (resp. increase of 0.01% (p<0.005). For any fixed values of the covariates in the model, including time, Cluster $[0, 3]$ have the highest percentage of cases (Table 5.6). Estimates for the main effects and the interaction effects are presented in Table 5.7.

Table 5.6: Estimates from the regression model for the relative AIDS incidence (within the HIV population) cross-classified by year of diagnosis and reporting delay quarter

Variable	Estimated Coefficient	Estimated Standard Error	t-value	P-value
Intercept	0.461	< 0.001	7.82E+08	< 0.001
Year of diagnosis	0.0104	< 0.001	1.09E+08	< 0.001
C2 ($[0, 3]$ is the reference class)	-0.415	0.0135	9.47E+02	< 0.001
Year of diagnosis * C2	-0.0103	< 0.001	1.43e+04	< 0.001

The homocedasticity and normality of the residuals were studied by graphical analysis and these assumptions did not seem to be compromised.

Individual delay curves can be obtained from (5.32) and these models can be used for estimating the expected percentage of cases for each delay.

Table 5.7: Individual models for the annual AIDS percentage cross-classified by diagnosis year and reporting delay quarter

Reporting delay	Estimated model
$(0, 3]$	$X_i = 0.461 + 0.0104i$
> 3	$X_i = 0.046 + 0.0001i$

5.2.3 Non-Parametric Estimates

In order to identify the main factors influencing the reporting delays in the HIV-AIDS cases within the Portuguese Surveillance System, several data mining models were considered: multilayer artificial neural networks (MLP), naive Bayesian classifiers (NB), support vector machines (SVM) and the K-Nearest Neighbour algorithm (KNN).

Data Pre-Processing As most HIV-AIDS cases were seen to be reported within the first 3 months after diagnosis, and the reporting behaviour seemed to be homogeneous from that onwards (Figure 5.20), reporting delays were discretized into two classes, with a cut-point at 3 months after diagnosis [22, 204]. This method may some of the administrative inaccuracies inherent to the continuous reporting delays. Patients nationalities were classified according to the World Health Organization (WHO) HIV-AIDS Regions. All residents in Portugal have access to health care services provided by the National Health Service (NHS). It is managed at 5 regional levels through Regional Health Administrators (RHA) that are accountable to the Ministry of Health: North, Centre, Lisbon and Vale do Tejo , Alentejo and the Algarve. Each RHA is responsible for the strategic management of its population health, supervision and control of hospitals, management of primary care/NHS primary care centres, centres for treatment of addictive behaviours, and implementation of the national health policy objectives [109]. The Ministry of Health cooperates with the Ministry of Justice for providing health care services on prisons and with the Ministry of Defence for providing health care services to the servicemen. Given this structure, the information on the health providers responsible for the diagnoses and reporting processes was cross-classified by type of health care institute and regional administration. Prisons and Military Institutions were considered a single type of health care provider. An additional category ('Admin') was created to accommodate observations without a specific health care provider. Supervised classification was then built upon that 2-class discretization, through feed-forward multilayer perceptron, naive networks, support vector machines and the K-nearest neighbour

algorithm with the following input features: age at diagnosis, gender, patient nationality, disease stage, HIV risk group, type of health provider, and administrative and financial responsibility of health care providers. To predict whether a case will ever be reported and to identify which factors influence that characteristic, we used feed-foward multilayer perceptron, naive neural networks, SVMs and the k-nearest neighbour algorithm.

Feed-foward multilayer perceptron

Artificial neural networks are a popular alternative to conventional statistical models [186]. In particular feedfoward multilayer perceptron networks with back-propagation training algorithms are the most widely used. They are effective in the analysis of complex data with non-linear trends and time-dependent covariates, and even high-order interactions [187].

In this section it was implemented a feed-forward multilayer perceptron with a hidden layer (with 18 neurons). Sigmoid functions were used as transfer and activations functions. For training, a back propagation algorithm with 500 training cycles, a learning rate of 0.3, a momentum of 0.2, and an error of $\epsilon = 1.0E^{-5}$ was used. The constants were finely tuned according to the data and results.

Naive bayesian classifier

A naive (or simple) Bayesian classifier (NB) is a probabilistic classifier which assumes that all attributes contribute equally, and independently, to the final decision [173]. It is a computational simple algorithm that can handle a data set with many attributes and thus widely-used in medical data mining.

Support Vector Machines Learning

Support Vector Machines (SVM) are amongst the most popular and efficient classification and regression methods currently available. These algorithms apply simple linear methods in a high-dimensional feature space that is non-linearly related to the input space. Usually, all attributes are employed and non-overlapping partitions are generated. In this thesis we used a SVM for classification purposes, with a sigmoid kernel of degree 3, a gamma parameter equal to the inverse of the sample size, a constant of the regularization term in the Lagrange formulation equal to one, a tolerance equal to 0.001, an error of $\epsilon = 0.1$ and a heuristic shrinking. The training set was randomly chosen and contained 80% of the initial data. The constants were finely tuned according to the data and results.

K - Nearest Neighbour

The k-nearest neighbour algorithm is one of the most popular classification algorithms. The algorithm performs a case partition in a pre-user-defined number of clusters by comparing a given test sample with a training sample. Each object is assigned to the class corresponding to the majority vote from its K nearest neighbours. “Closeness” is defined in terms of a distance metric. For the given data, the K-means algorithm with 2 clusters was applied, with a mixed euclidean distance given the nominal and the quantitative nature of the input features.

A 10-fold cross-validation was used for MLP, NB and KNN validation. A leave-one-out method was applied to the SVM model.

Table 5.8 describes the performance of the data mining algorithms. The accuracy in predicting the class membership ranged from 53% (NB) to 63% (MLP) approximately, the precision ranged from 16% (NB) to 76% (MLP) and finally the recall ranged from 60% (MLP,

NB and SVM) to 66%(KNN). Excluding the SMV which had a different fitting process, the fastest algorithm needed 15s (Naive) to produce the results and the slowest approximately 51 min (MLP).

Table 5.8: MLP, KNN, NAIVE and SVM performance

	Confusion matrix			Accuracy	Precision	Recall	Time
	Actual: 0-3	Actual: >3	sum (%)				
MLP							
Predicted: 0-3	13831	9136	22967 (60%)	62.98% +/-	76%	60%	50min37s
Predicted: >3	4381	9162	13543 (68%)	0.78%			
sum (%)	18212 (76%)	18298 (50%)					
KNN							
Predicted: 0-3	8344	4261	12605 (66%)	61.30% +/-	45%	66%	2 min
Predicted: >3	9868	14037	23905 (59%)	0.88%			
sum (%)	18212 (45%)	18298 (77%)					
Naive							
Predicted: 0-3	2997	1983	4980 (60%)	52.90% +/-	16%	60%	15s
Predicted: >3	15215	16315	31530 (48%)	0.97%			
sum (%)	18212 (16%)	18298 (11%)					
SVM							
Predicted: 0-3	12795	8412	21207 (60%)	62%	70%	60%	2min
Predicted: >3	5417	9886	15303 (65%)				
sum	18212 (70%)	18298 (54%)					

Analysing the behaviour of these two groups according to the year of diagnosis, it can be seen that it is mostly constant until the most recent years (in the last years the percentage of cases is biased due to the reporting delay). Moreover, the two groups seemed to behave similarly with respect to the patient's age at the diagnosis, gender, stage of the disease, transmission risk group, nationality, type of the health care provider that made the diagnosis and administrative and financial responsible from the health care provider (Table 5.1).

Tables 5.8 show that MLP provided the best results, with a higher classification accuracy (approximately 63%), precision (approximately 76%) and recall (approximately 60%), but on the other hand it was considerably slower. It can predict, with reasonable efficiency the group of reporting delays less than 3 months long. The SVM model provided similar results and was considerably faster.

While around 60% of the accuracy may be a reasonable result, it can be explained by characteristics of the input data. In many cases,the quality of the data within the biomedical and healthcare fields is inferior to that found in other fields [173]. In our data set the main reasons for the poor classification quality are most probably related to stigma around the disease that leads patients to provide incorrect informations (mainly with respect to the transmission group), to high demands of the healthcare systems and to the implementation of the surveillance system, more specifically paper form reports and poor communication between the stakeholders. In a previous qualitative assessment of the Portuguese Surveillance System, Mauch pointed out that all clinicians reported that they complete the notification form after the patient has left the office, sometimes several days or weeks later. This practice has the potential to contribute to inaccuracies in reporting for some variables, such as the associated risk group, due to recall errors[93]. Moreover, errors may also arise from the transcription of the information in the paper report to databases.

5.3 Continuous Outcome - Multilevel Regression

Since almost cases were reported within the reporting delay $[0, 3]$ months, we decided to model delays in continuous time and using a fully parametric approach. The parametric survival analysis is an appropriate method because reporting delay can be viewed as a 'time from diagnosis to report'. This approach avoids the need to, somehow arbitrarily,

classify the delays into discrete categories which considers that no matter whether a person is diagnosed at the beginning or at the end of a time unit, a same reporting delay proportion applies [119] . When the time unit is large, such as a quarter, this can yield very imprecise estimates for recent incidence. Moreover, by fitting a continuous long tail distribution we are allowing the model to assess the underreporting, since this issue can be considered as a particular case of reporting delay with infinite length.

Similarly to Noufailly, using a continuous time framework the reporting delay distribution can span over all values helping rinse a better understanding of the processes generating the delays [198]. Finally, we decided to avoid the reverse time proportional hazards regression model in favour of an explicit model for the reporting delay (measured in number of days), using common distributions in survival analysis parametric models: normal, log-normal, exponential, Weibull and gamma. This allows a flexible representation of the reporting delay distribution at moderately long delays and accommodates the observed extremely long tails, which are a striking feature of these data [198]. Moreover, if the cases are reported in batches, the multinomial assumption does not capture the patterns and if the discretized time intervals are large this can yield very imprecise estimates for recent HIV-AIDS incidence.

Since the data is only right - truncated (there are no censored reporting delay length), the inferences on covariate effects are mainly driven by the shorter delays, which are much more numerous than longer delays [198]. This is appropriate for rapid detection of changes in epidemic patterns.

Moreover, since reporting delay is an administrative process, and considering the organizational structure of the National Health System and historical changes on the HIV -AIDS Surveillance System, we considered fitting multilevel regression models with a distribution consistent with the parametric survival analysis approach.

So, the main aim is to describe the reporting delay distribution gaining a better understanding of the reporting process taking into consideration the individual, administrative, organizational and historical context, and to investigate whether individual and structural factors influence the reporting delays. Cases diagnosed more recently will be weighted more heavily than those diagnosed further in the past, reflecting the fact that less of the recently diagnosed cases have been reported.

For fitting purposes, we follow the strategy from simple to complex models following the next steps for the normal, log-normal, gamma and Weibull:

1. Fit a simple generalized linear model, without a multilevel structure, for validating the covariates and the choice of the appropriate distribution. Selection of the best model is based on the Bayesian Information Criterion (BIC);

2. Fit the null multilevel (containing no covariates) model for validating the multilevel structure;

3. Fit a multilevel model adding the covariates found in 1.

4. Fit a multilevel model adding higher-level covariates.

5.3.1 The specification of the reporting delay model

In this preliminary step we investigate whether if the reporting delays should be modelled directly or using a logarithm transformation (AFT) and which distribution should be specified. This method presents several advantages when compared to the traditional Proportional Hazards (PH) model, namely:

- it does not require the assumption of proportional hazards;

- it models directly the effect of covariates on the survival, so the interpretation of the results is clearer and easier (in terms of effects on the mean survival time) than the PH model, where we model the effect of covariates on a conditional probability;
- it is simpler to evaluate and it can be extended to more complex situations such as, for example, interactions between covariates and time, or include non-parametric effects, of some covariates or covariates with time-dependent parameters [217].

The following covariates were taken into consideration:

- Age of the patient (continuous)
- Notification Year (continuous)
- Nationality classified using WHO system for HIV-AIDS epidemic (European, African - reference level and American)
- Sex (female as the reference level)
- Disease stage (AIDS and Asymptomatic, being the last the reference level)
- Risk-group (Heteressexual - reference level, IDU and MSM)
- Type of health service (Admin meaning not specific identification of type of service but its known the region were the report took place, Hospital, Primary Care, Health services shared with Justice Ministery and with Defence Ministery, Other/Unknown and Addictive Behaviours - reference level)
- Administration Region (RHA North , LVT, Algarve, Alentejo - reference level and Centro and the same for IDT's which, until the end of the data collection period, were independent from RHA's)
- Recent incidence of the institution determined by the number of cases diagnosed in the same quarter as the notified case; the idea here is to use the local incidence as a proxy for the notifications' overload.
- Historical epocs (time periods with specific surveillance rules and regulations defined by special alterations of the surveillance system or introduction of a new treatment or prevention actions: E2 - from 1992 until 1997 -reference level, E3 - from 1997 until 2005, E4 - from 2005 until 2009 and finally E5 - from 2009 until the present days.)

We considered that the data collected before 1992 had not enough quality for performing any regression model, so the data were truncated at this time stamp.

Models with the above covariates and errors following the normal, log-normal, Weibull and Gamma distribution (with log-link for consistency with the other regression models). The estimates of the coefficients, their standard errors as well as model BIC criterion are presented on (Table 5.9). All models were fitted using the iteratively reweighted least squares (IWLS) method.

Comparing the Weibull and the Gamma models, it can be seen that with some point exceptions, the coefficients are remarkably similar to each other. Even the standard errors are almost identical. Both have very close coefficient estimates and standard errors to those from the log-normal distribution. These models are not nested since they assume different distributions for the response, which makes direct comparison problematic. Note that purely numerical comparisons such as AIC and BIC (since these criterium uses the likelihood function which depends of very different distributions) are risky and that some attention to residual diagnostics, scientific context and interpretation is necessary [168].

Table 5.9: The estimates and standard error of a full model for the normal, log-normal, Weibull and gamma distribution

	Normal	Log-Normal	Weibull	Gamma
	Estimate (Std. Error)	Estimate (Std. Error)	Estimate (Std. Error)	Estimate (Std. Error)
Intercept	-9550 (2025.00)	-23.880 (8.50)	-39.785 (8.42)	-41.880 (7.42)
Age	-0.803 (0.16)	-0.003 (0.001)	-0.004 (0.001)	-0.004 (0.001)
Notification Year	4.900 (1.01)	0.014 (0.004)	0.022 (0.004)	0.024 (0.004)
Male	-1.693 (4.22)	-0.021 (0.018)	-0.016 (0.017)	-0.016 (0.015)
AIDS	-83.010 (3.68)	-0.448 (0.015)	-0.345 (0.014)	- 0.332 (0.013)
America	-6.894 (12.32)	-0.013 (0.052)	-0.016 (0.049)	-0.014 (0.045)
Europe	9.967 (6.07)	0.059 (0.025)	0.033 (0.024)	0.031 (0.022)
IDU	57.440 (4.59)	0.193 (0.019)	0.222 (0.018)	0.220 (0.017)
MSM	-7.950 (5.87)	0.011 (0.025)	-0.034 (0.024)	-0.038 (0.022)
Admin	-0.470 (19)	0.496 (0.080)	-0.031 (0.077)	-0.062 (0.070)
Hospital	1.322 (18.46)	0.433 (0.077)	-0.048 (0.075)	-0.074 (0.068)
INST. / SERV. - MD	17.900 (34.93)	0.465 (0.147)	-0.046 (0.141)	-0.072 (0.128)
INST. / SERV. - MJ	-118.100 (21.05)	-0.039 (0.088)	-0.620 (0.085)	-0.646 (0.077)
Other/Unknown	92.440 (46.45)	0.666 (0.195)	0.268 (0.187)	0.239 (0.170)
Primary care	-56.050 (20.08)	0.095 (0.084)	-0.312 (0.081)	-0.332 (0.074)
RHA ALGARVE	-211.100 (17.52)	-0.856 (0.074)	-0.727 (0.071)	-0.705 (0.064)
RHA CENTRO	-220.30 (16.66)	-0.814 (0.070)	-0.744 (0.066)	-0.726 (0.061)
RHA LVT	-135.80 (15.86)	-0.521 (0.067)	-0.431 (0.064)	-0.418 (0.058)
RHA NORTE	-196.70 (16.15)	-0.750 (0.068)	-0.668 (0.065)	-0.652 (0.059)
IDT ALENTEJO	-356.20 (49.83)	-1.031 (0.209)	-1.579 (0.201)	-1.602 (0.183)
IDT ALGARVE	-348.70 (33.88)	-1.178 (0.142)	-1.547 (0.137)	-1.541 (0.124)
IDT CENTRO	-218.20 (52.02)	-0.725 (0.218)	-0.816 (0.21)	-0.811 (0.191)
IDT LVT	-193.80 (31.93)	-0.493 (0.134)	-0.660 (0.129)	-0.661 (0.117)
IDT NORTE	-346.30 (40.65)	-1.301 (0.171)	-1.567 (0.164)	-1.554 (0.149)
E3	122.100 (12.01)	0.684 (0.050)	0.915 (0.049)	0.935 (0.044)
E4	94.720 (15.30)	0.593 (0.064)	0.832 (0.063)	0.850 (0.056)
E5	36.550 (17.94)	0.314 (0.075)	0.542 (0.074)	0.563 (0.066)
Recent Incidence	0.429 (0.03)	0.002 (0.000)	0.002 (0.0001)	0.002 (0.000)
AIC	371977.1	82282.67	337156.00	337894.8
BIC	372214.4	82519.99		338132.1

For a better comparing the estimates we consider the plot of the estimates together with their profile confidence interval, for each regression model (Figures 5.40 until 5.43) with the exception of the Intercept due to its large coefficient. From these plots, although with different scales, the estimates provided by the gamma distribution are similar to the ones given by the normal distribution. From the comparison between the estimated obtained from normal and the log-normal models, it can be seen that the former yielded narrower confidence intervals and from the comparison between the log-normal and gamma it can be seen that some coefficients changed statistical significance, such as 'Primary care', 'Other / unknown', 'Inst. / Serv. MJ', 'Inst. / Serv. MD' , 'Hospital' and 'Admin' all from the same covariate which represents the type of health provider.

The skewness of the confidence intervals obtained from the Weibull model allowed us to restrict the residuals' analysis to the normal, log-normal and gamma models. (Figures 5.44, 5.45 and 5.46).

The plots of the residuals against the fitted values show a linear decreasing pattern. This pattern is often associated with counting variables and when there are lots of observations with the same values. These hard limits are also responsible for the apparent set of points lying on a straight line in the scale-location plots. Moreover, there is a suggestion of a mild association between residuals and fitted values (Figures 5.44, 5.45 and 5.46). Another important characteristic is that the residuals have approximately the same variance,

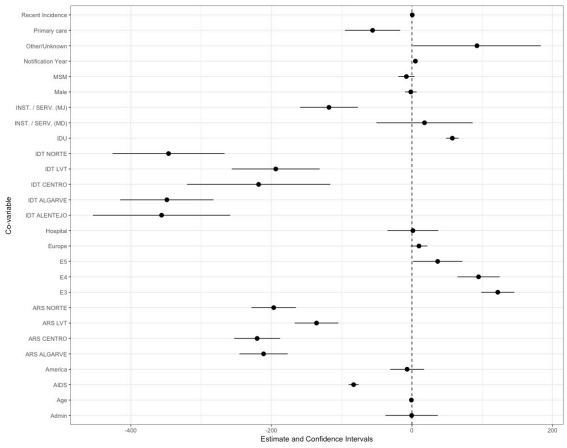


Figure 5.40: The estimates and profile confidence intervals from the normal distribution

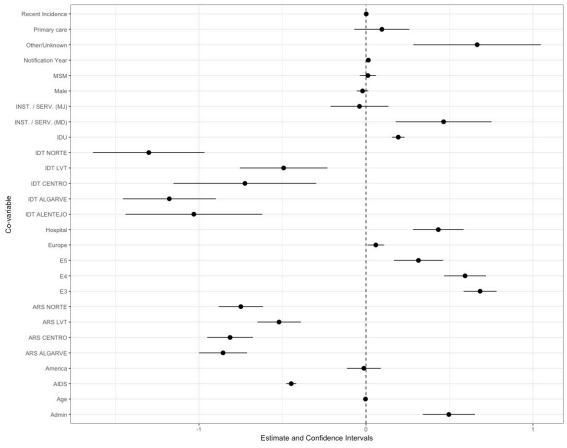


Figure 5.41: The estimates and profile confidence intervals from the log normal distribution

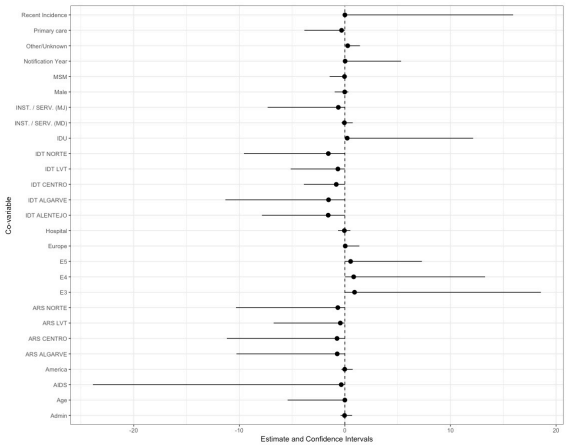


Figure 5.42: The estimates and profile confidence intervals from the Weibul distribution

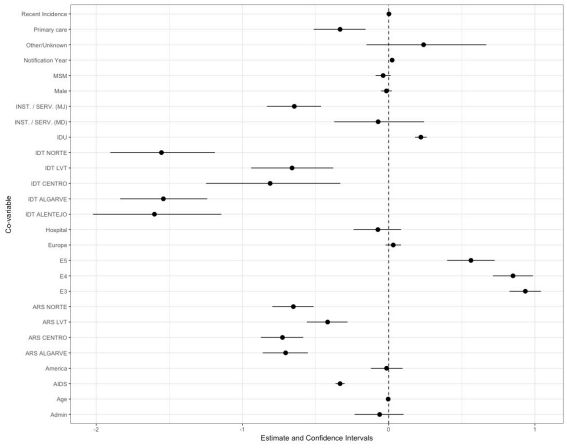


Figure 5.43: The estimates and profile confidence intervals from the gamma distribution

regardless of the predicted values which is a suggestion of homocedasticity. The normal quantile-quantile (QQ) plot do not reveal serious problems on the adherence to the normal distribution (Figures 5.44, 5.45 and 5.46).

The collection of the previous findings led to the choice of the log-normal distribution for the response. The stepwise algorithm based on AIC was then applied as a variable selection method. The obtained model is presented in Table 5.10.

The variance of Y depends on both the expected values of Y, μ_y, and the variance of $Z = ln\,(Y)$, σ_Z^2; $Var\,(Y|X) = \sigma_{Y|X}^2 = \left(exp\left(\sigma_z^2\right) - 1\right)\mu_y^2$

Table 5.10: The estimates of the log-normal model

	Estimate	EXP(Estimate)	Std. Error	t value	Pr(>|t|)
Intercept	-333.80	0.000	7.22	-46.22	< 0.001
Age	-0.003	0.997	0.00	-4.009	< 0.001
Notification Year	0.17	1.18	0.00	46.961	< 0.001
AIDS	-0.45	0.64	0.01	-30.708	< 0.001
America	0.05	1.06	0.05	1.113	0.266

Table continues...

Table 5.10: The estimates of the log-normal model

	Estimate	EXP(Estimate)	Std. Error	t value	Pr(>|t|)
Europe	0.08	1.08	0.02	3.293	0.001
IDU	0.20	1.22	0.02	11.394	< 0.001
MSM/Bisexual	0.03	1.03	0.02	1.249	0.212
Admin	0.49	1.64	0.08	6.52	< 0.001
Hospital	0.39	1.47	0.07	5.261	< 0.001
INST. / SERV. - MD	0.30	1.35	0.14	2.149	0.032
INST. / SERV. - MJ	0.00	1.002	0.08	0.029	0.977
Other/Unknown	0.49	1.624	0.19	2.62	0.009
Primary care	0.10	1.109	0.08	1.289	0.198
ARS ALGARVE	-0.79	0.455	0.07	-11.261	< 0.001
ARS CENTRO	-0.74	0.477	0.07	-11.118	< 0.001
ARS LVT	-0.44	0.642	0.06	-7.004	< 0.001
ARS NORTE	-0.66	0.516	0.06	-10.253	< 0.001
IDT ALENTEJO	-0.99	0.370	0.20	-4.993	< 0.001
IDT ALGARVE	-0.99	0.373	0.14	-7.287	< 0.001
IDT CENTRO	-0.82	0.438	0.21	-3.972	< 0.001
IDT LVT	-0.52	0.596	0.13	-4.058	< 0.001
IDT NORTE	-1.21	0.298	0.16	-7.451	< 0.001
E3	-0.79	0.452	0.04	-21.963	< 0.001
E4	-1.96	0.141	0.05	-39.788	< 0.001
E5	-2.85	0.058	0.06	-48.525	< 0.001
Recent Incidence	0.001	1.001	0.00	13.132	< 0.001

The residuals of this model were inspected and the model assumptions did not seem to be compromised.

From the model and assuming that all other values remain constant, it can expect that with an increase of year of the patient age, we would expect an decrease of 0.3% of the reporting delay. An increase of one notification year will produce approximately 18% of change in the reporting delay and one case increase of the incidence will produce and change of 0.1% in the reporting delay. An AIDS case is approximately 36% faster than an asymptomatic case and an American case is almost 6% slower than an African. An European case is approximately 8% slower than an African case. A IDU case is 20% slower than a Heterossexual and a MSM / Bisexual is almost 3% slower than a Heterossexual case. A reported case from an Administrative entity is almost 64% slower than an entity of type 'Addictive Behaviours'. An Hospital is 47% slower than a type 'Addictive Behaviours' entity. The shared services between the Defense Ministry and the Health is 34% slower than type 'Addictive Behaviours' and the shared services with the Ministry of Justice is only 0.23% slower than type 'Addictive Behaviours'. Other/ Unknown is 62% slower than type 'Addictive Behaviours'. Primary Care is 11% slower than type 'Addictive Behaviours'. The RHA of Algarve is 54% faster than RHA Alentejo. The RHA LVT is 35% faster than RHA Alentejo. IDT Alentejo is 63% faster than RHA Alentejo. IDT Algarve is 62% faster that RHA Alentejo. IDT Centro is 56% faster than RHA Alentejo. IDT LVT is 40% faster than RHA Alentejo. IDT Norte is 70% faster than RHA Alentejo. The cases reported in E3 were 54% faster than E2; the cases reported in E4 were 85% faster than in E2; the cases reported in E5 were 94% faster than E2.

From the model above it can be seen that the institutional variables such as: Type of services and Region of Health Administration (RHA) have high variability and high standard deviations suggesting a multilevel structure underlying the model.

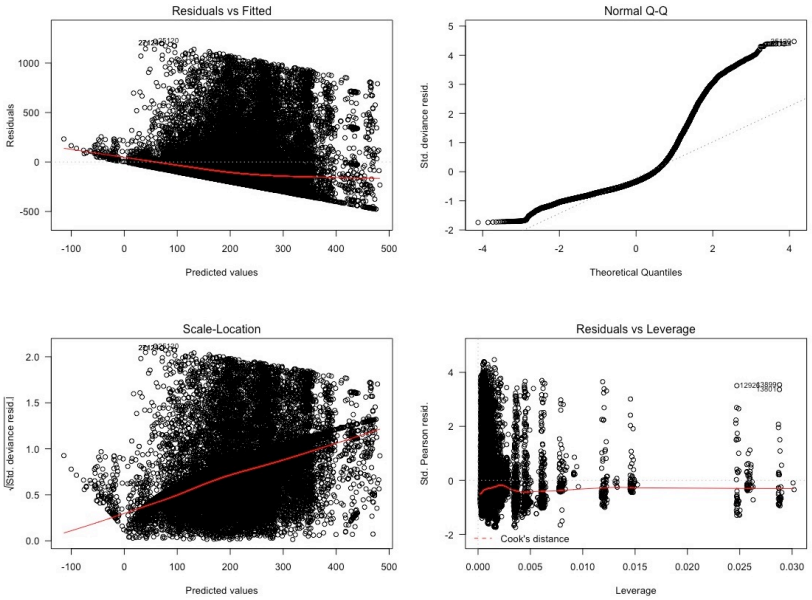


Figure 5.44: Residuals from normal distribution

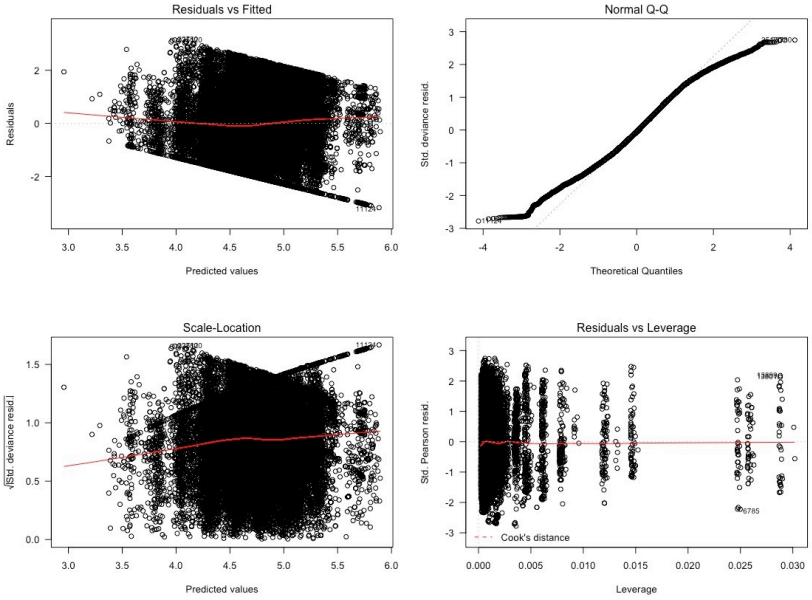


Figure 5.45: Residuals from log normal distribution

5.3.2 Reporting delay with a Multilevel Structure

In order to determine the best multilevel structure, a full theoretical model was considered. The first approach was developed based on the contextual framework of this problem. Namely, individuals (Population) are diagnosed in a health service institution that gets the case reported (Reporting Entity), which in turn belongs to a certain type of service, such as Hospital care, primary care, additive behaviours... (Type of Service). These entities are under a Region of Health Administration, which in turn follows therapeutic, administrative rules and regulations defined by Scientific Communities and Governments. This framework is represented on Figure 5.47.

This context can be expressed as the following multilevel structure

Level 1 - Population (i=1,..., 27461)

$$log(Y_{ijklm}) = \beta_{0jklm} + \epsilon_{ijklm}$$

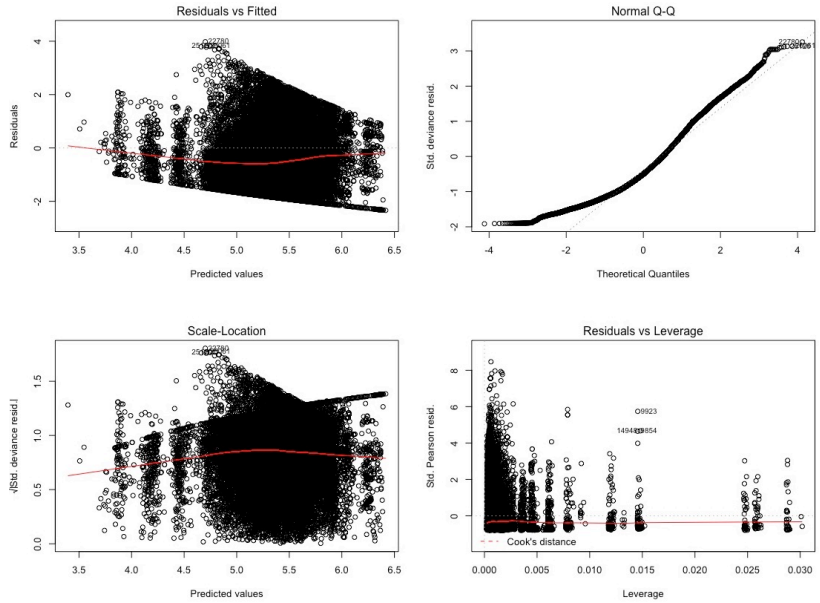


Figure 5.46: Residuals from log gamma distribution

Figure 5.47: Conceptual representation of the nested structure

where Y_{ijklm} is the reporting delay measured in days.

Level 2 - Reporting Entity -Group Level (j=1,..., 593)

$$\beta_{0jklm} = \gamma_{00klm} + \nu_{jklm}$$

Level 3 - Type of service - Group Level(k=1,..., 100)

$$\gamma_{00klm} = \theta_{000lm} + \upsilon_{klm}$$

Level 4 - Administrative Responsibility - Group Level (l=1,...,38)

$$\theta_{000lm} = \psi_{0000m} + \varpi_{lm}$$

Level 5 - Historical and Context Epocs - Group Level (m=1,..., 4)

$$\psi_{0000m} = \varsigma_{00000} + \varphi_m$$

where,

$$\epsilon_{ijklm} \sim N\left(0, \sigma^2\right)$$
$$\nu_{jklm} \sim N\left(0, \sigma_1^2\right)$$
$$\upsilon_{klm} \sim N\left(0, \sigma_2^2\right)$$
$$\varpi_{lm} \sim N\left(0, \sigma_3^2\right)$$
$$\varphi_m \sim N\left(0, \sigma_4^2\right)$$

The model was fitted with the restricted maximum likelihood (REML) and the results are represented on Table 5.11.

Table 5.11: Null Multilevel Model

Groups	Variance	Std.Dev.	Std. Dev. 2.5 % C.I. Limit	Std. Dev. 97.5 % C.I. Limit
Random Effects				
Level 2 - σ_1	0.36	0.6	0.45	0.56
Level 3 - σ_2	0.25	0.50	0.30	0.55
Level 4 - σ_3	0.002	0.042	0.00	0.32
Level 5 - σ_4	0.17	0.41	0.13	0.80
Residual - σ	1.18	1.09	0.93	0.94
Fixed Effects	Estimate	Std. Error		
Intercept	4.68	0.1636		

Note that, σ is much higher than the other standard deviations and there is evidence that σ_3 and σ_4 are not significant due to the curvilinear form of Figure 5.48.

The profile zeta plot ζ for the fixed - effect parameter (Intercept) term is slightly over-dispersed relative to the normal distribution. While for σ, σ_1 and σ_2 have a good normal approximation while the standard deviations for σ_3 and σ_4, are skewed. The skewness for σ_4 is worse than that for σ_3, making the estimate of σ_4 less precise than of σ_3 , in both absolute and relative senses (Figure 5.48). For an absolute comparison we compare the widths of the confidence intervals ($\sigma_3 \in [0.00; 0.32]$ and $\sigma_4 \in [0.13; 0.80]$) for these parameters. Clearly, the lack of precision of these estimates is a consequence of only having 4 distinct levels of the Epocs factor.

In general, it is more difficult to estimate a measure of spread, such as the standard deviation, than to estimate a measure of location, such as a mean, especially when the number of levels of the factor is small. Six levels are about the minimum number required for obtaining sensible estimates of standard deviations for simple, scalar random effects terms [179].

So, we proceed to fit several null nested models in order to validate the minimal multi-level structure.

Model A - random effect on the Reporting Entity
Level 1 - Population (i=1,..., 27461)

$$log(Y_{ij}) = \beta_{0j} + \epsilon_{ij} \tag{5.33}$$

Level 2 - Reporting Entity - Group Level (j = 1,..., 351)

$$\beta_{0j} = \gamma_{00} + \nu_j \tag{5.34}$$

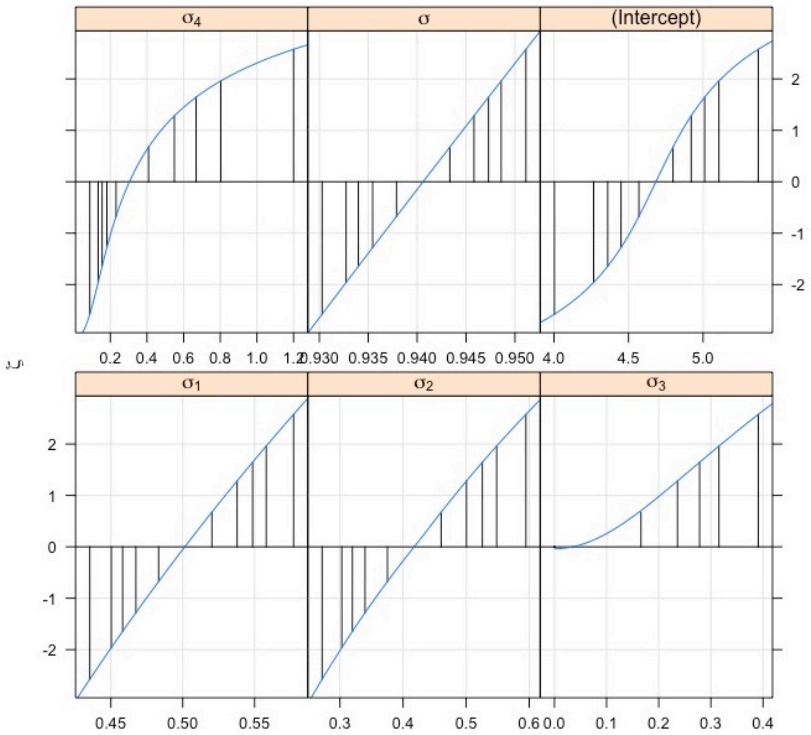


Figure 5.48: Profile zeta plot of the parameters in full null hierarchical model.

where,

$$\epsilon_{ij} \sim N\left(0, \sigma^2\right)$$
$$\nu_j \sim N\left(0, \sigma_1^2\right)$$

Model B - random effect by Entity nested within the Type of Service:
Level 1 - Population (i=1,..., 27461)

$$log(Y_{ijk}) = \beta_{0jk} + \epsilon_{ijk} \tag{5.35}$$

Level 2 - Reporting Entity - Group Level (j = 1,..., 351)

$$\beta_{0jk} = \gamma_{00k} + \nu_{jk} \tag{5.36}$$

Level 3 - Type of Service - Group Level (k = 1,..., 8)

$$\gamma_{00k} = \theta_{000} + \upsilon_k \tag{5.37}$$

where,

$$\epsilon_{ijk} \sim N\left(0, \sigma^2\right)$$
$$\nu_{jk} \sim N\left(0, \sigma_1^2\right)$$
$$\upsilon_k \sim N\left(0, \sigma_2^2\right)$$

Model C - random effect by Reporting Entity, nested within the Type of Service which are nested in Administrative Responsibility Level 1 - Population (i=1,..., 27461)

$$log(Y_{ijkl}) = \beta_{0jkl} + \epsilon_{ijkl} \tag{5.38}$$

Level 2 - Reporting Entity - Group Level (j = 1,..., 354)

$$\beta_{0jkl} = \gamma_{00kl} + \nu_{jkl} \tag{5.39}$$

Level 3 - Type of Service - Group Level (k = 1,..., 30)

$$\gamma_{00kl} = \theta_{000l} + \upsilon_{kl} \tag{5.40}$$

Level 4 - Administrative Responsibility - Group Level (l = 1,..., 10)

$$\theta_{000l} = \psi_{0000} + \varpi_{l} \tag{5.41}$$

where,

$$\epsilon_{ijkl} \sim N\left(0, \sigma^2\right)$$
$$\nu_{jkl} \sim N\left(0, \sigma_1^2\right)$$
$$\upsilon_{kl} \sim N\left(0, \sigma_2^2\right)$$
$$\varpi_{l} \sim N\left(0, \sigma_3^2\right)$$

Model D: random effect by Reporting Entity, nested within the Type of Service which are nested in Administrative Responsibility, nested within Historical Rules and Regulations

Level 1 - Population (i=1,..., 27461)

$$log(Y_{ijklm}) = \beta_{0jklm} + \epsilon_{ijklm}$$

where Y_{ijklm} is the reporting delay measured in days.

Level 2 - Reporting Entity -Group Level (j=1,..., 593)

$$\beta_{0jklm} = \gamma_{00klm} + \nu_{jklm}$$

Level 3 - Type of service - Group Level(k=1,..., 100)

$$\gamma_{00klm} = \theta_{000lm} + \upsilon_{klm}$$

Level 4 - Administrative Responsibility - Group Level (l=1,...,38)

$$\theta_{000lm} = \psi_{0000m} + \varpi_{lm}$$

Level 5 - Historical and Context Epocs - Group Level (m=1,..., 4)

$$\psi_{0000m} = \varsigma_{00000} + \varphi_{m}$$

where,

$$\epsilon_{ijklm} \sim N\left(0, \sigma^2\right)$$
$$\nu_{jklm} \sim N\left(0, \sigma_1^2\right)$$
$$\upsilon_{klm} \sim N\left(0, \sigma_2^2\right)$$
$$\varpi_{lm} \sim N\left(0, \sigma_3^2\right)$$
$$\varphi_{m} \sim N\left(0, \sigma_4^2\right)$$

The interclass correlation associated to the group level Reporting Entity of model A is approximately equal to 29.67%. Thus, 29.67% of the variance of the reporting delay is at the Reporting Entity group level. Considering the model B, 20.24% of the variance of the reporting delay is at the Reporting Entity and 8% is at the Type of Services. For the model C, 18.37% of the variance of the reporting delay is at the Reporting Entity, 10.50% is at the Type of services group level and 1.42% is at the Administrative Responsibility group level. For the model D, 18.15% of the variance is at the Reporting Entity level, 12.89% at the Type of Services level, 0.09% at the Administrative Responsibility level and 8.66% at the Historical Rules and Regulations level (Table 5.12). Since these models are intercept - only models that do not contains no explanatory variables, the residuals variances represent unexplained error variance.

Table 5.12: Null Models I.C.C.

	Model A		Model B		Model C		Model D	
	σ^2	I.C.C.	σ^2	I.C.C.	σ^2	I.C.C.	σ^2	I.C.C.
Reporting Entity	0.537	29.67%	0.359	20.24%	0.336	18.37%	0.357	18.15%
Type of Services			0.142	8.00%	0.192	10.50%	0.253	12.89%
Administrative Responsibility					0.026	1.42%	0.002	0.09%
Historical Rules and Regulations							0.170	8.66%
Residuals	1.273		1.273		1.273		1.183	

We found that the model B was the best based on the results from the Likelihood Ratio test. The tests statistics are presented on Table 5.13. It can be seen that B improve A, but C does not improve B. So, the multilevel structure represented by model B was select.

Table 5.13: Models Evaluation

Models:	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
A	3	80188	80213	-40091	80182			
A - B	4	77085	77118	-38538	77077	3105.7	1	<0.001
B - C	5	77100	77141	-38545	77090	0.001	1	1.000
C - D	6	75359	75409	-37674	75347	1742.5	1	<0.001

The profile zeta plot for the constant intercept (β_0) for the model B, which it was called the minimal null model has almost symmetric intervals. The profile zeta plot for σ, σ_1 appears to be a good normal approximation and although the plot σ_2 is not a straight line, the bias is very mild 5.49.

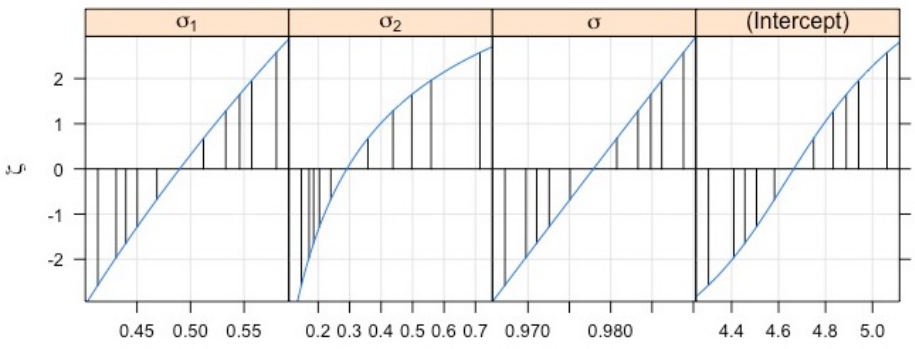


Figure 5.49: Profile zeta plot of the parameters in the minimal multilevel model

For the next experience we added the covariates according to the minimal set determine in (5.10) model. It was considered just a random intercept:

Level 1 Model

$$\begin{aligned} log(Y_{ijk}) &= \beta_{0jk} + \beta_1 DataNotification_{ijk} + \beta_2 RiskGroup_{ijk} + \\ &\quad \beta_3 Nationality_{ijk} + \beta_4 Stage_{ijk} + \beta_5 Age_{ijk} + \\ &\quad \beta_6 RecentIncidenceijk + \epsilon_{ijk}, \text{ with } \epsilon_{ijk} \sim N\left(0, \sigma^2\right) \end{aligned} \tag{5.42}$$

Level 2 Model

$$\beta_{0jk} = \gamma_{00k} + \psi_{0jk}, \text{ with } \psi_{0jk} \sim N\left(0, \sigma_1^2\right) \tag{5.43}$$

Level 3 Model

$$\gamma_{00k} = \theta_{000} + \nu_{00k}, \text{ with } \nu_{00k} \sim N\left(0, \sigma_2^2\right). \tag{5.44}$$

.

We used the restricted likelihood to fit the models and Bayesian approaches and found that there were no significant changes in the coefficient values. So, only the results obtained from REML are presented. From the model and assuming that all other values remain constant, it can expect that with an increase of one year in the notification date it would be expected an increase of 0.1% of the reporting delay. An increase of one year in patient age will produce approximately -0.31% of change in the reporting delay and one case increase in the recent incidence will produce an increase of 0.23% in the reporting delay. An AIDS case is approximately 31% faster than an asymptomatic case and an American case is almost 7% faster than an African. An European case is approximately 2.1% slower than an African case. A case from Low incidence country may be 22.5% slower that an African (note: there are just a hand full of this cases). A IDU case is 16% slower than a Heterossexual and a MSM / Bisexual is almost 3.6% faster than a Heterossexual case.

Note that the variables $Intercept$, $NotificationDate$ and $Nationality$ as well as the category $MSM/Bisexual$ are not statistically significant. The Intra Class Correlation Coefficient (ICC) for the Type of Services level is 10% and for the Reporting Entities level is 21% (Table 5.14).

Table 5.14: Model with a random intercept and population level covariates

		Estimate	Std. Error	95% I.C	
Fixed					
Intercept		2.714	3.290	-3.740	9.170
Notification date		0.001	0.002	-0.002	0.004
Risk Group	IDU	0.150	0.015	0.120	0.180
	MSM/Bisexual	-0.036	0.019	-0.070	0.001
Nationality	America	-0.070	0.040	-0.150	0.010
	Europe	0.022	0.021	-0.020	0.060
	Low Incidence Countries	0.200	0.160	-0.110	0.520
Stage	AIDS	-0.310	0.012	-0.330	-0.280
Age		-0.003	0.0006	-0.004	-0.002
Recent incidence		0.002	0.0001	0.002	0.003
Random					
Level 3 (Between Type of Services, 8 levels)	std	0.36			
Level 2 (Between Entities, 349 levels)	std	0.5			

(Table continues...)

Table 5.14: Model with a random intercept and population level covariates - continue

		Estimate	Std. Error	95% I.C
Level 1				
(Between individuals)	std	0.9		
Number of obs:	26461			

Analysing the profile zeta plot no significant problem was detected (Figure 5.50).

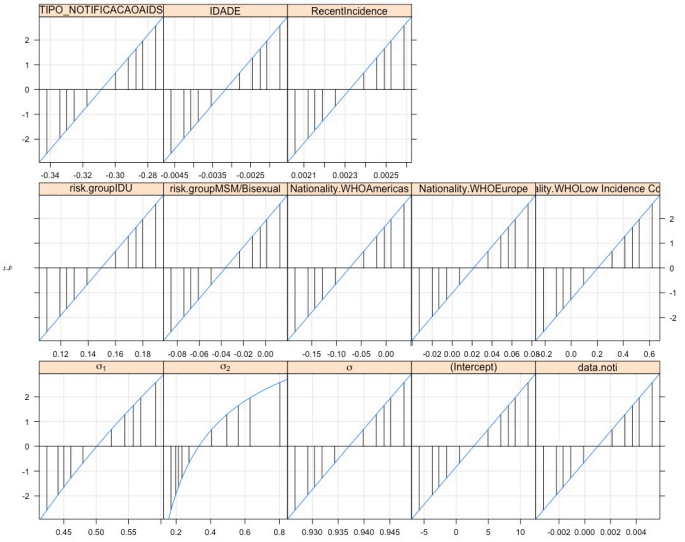


Figure 5.50: Profile zeta plot of the parameters of the random intercept model.

The prediction intervals for the random effects (Figures 5.51 (each line of the chart corresponds to a Health Reporting Institution, and since there are 349 levels the names were omitted for clarity representation) and 5.52) confirm that the conditional distribution of the random effects for type of services has a lower variability than does the conditional distribution of the random effects for the Reporting Entity, in the sense that the dots in Figure 5.52 have a lower variability than those in Figure 5.51. (Note the different independent axes for the two figures.) However, the conditional distribution of the random effects for some particular Type of services can have a lower variability than the conditional distribution of the random effects of particular reporting entities. Also, Hospitals seem to take longer to report a case than the Primary care services.

The final experiment consisted of adding a random slope allowing for the possibility that, for example, Entities with a higher reporting delay may, on average, be more strongly affected by recent incidence registered on its services. This model, which incorporates a random intercept and a random ”slope” for Recent Incidence, can be written as:

Level 1 Model

$$\begin{aligned} log(Y_{ijk}) &= \beta_{0jk} + \beta_1 DataNotification_{ijk} + \beta_2 RiskGroup_{ijk} + \\ &\quad \beta_3 Nationality_{ijk} + \beta_4 Stage_{ijk} + \beta_5 Age_{ijk} + \\ &\quad \beta_6 RecentIncidence_{ijk} + \epsilon_{ijk} \end{aligned} \tag{5.45}$$

Level 2 model

$$\beta_{0jk} = \gamma_{00k} + \gamma_{1jk} RecentIncidence_{0jk} + \psi_{0jk}, \text{ with } \psi_{0jk} \sim N\left(0, \sigma_1^2\right) \tag{5.46}$$

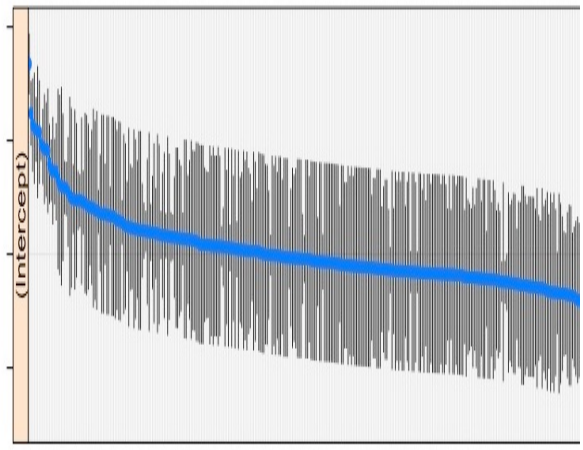


Figure 5.51: 95% prediction intervals for the random effect Entity

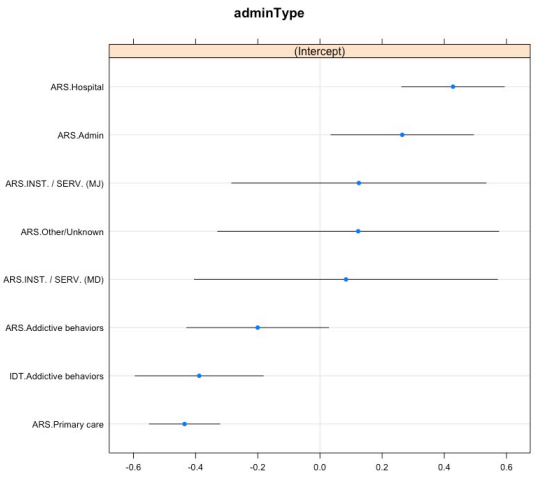


Figure 5.52: 95% prediction intervals for the random effect Administration

Level 3 model

$$\gamma_{00k} = \theta_{000} + \nu_{00k}, \text{ with } \nu_{00k} \sim N\left(0, \sigma_2^2\right) \tag{5.47}$$

and with ϵ_{ijk} following the a normal distribution.

We now interpret the statistically significant fixed-effects of the model. Assuming that the values from the remaining covariates are kept constant, it it is expected that: any 1-year increase in the notification date increases the reporting delay by 0.5%; any 1-year increase in the patients' age decreases the reporting delay by 0.3% and an 1-patient increase in the absolute incidence increases the reporting delay by 4%. Moreover, an AIDS case is expected to be approximately 25% faster to report than an asymptomatic case. An IDU case is expected to be 14% slower than a Heterosexual while no significant differences regarding the expected reporting delay were identified for MSM/Bisexuals and Heterosexuals. (Table 5.15).

Table 5.15: Estimates from the model with a random intercept and a random slope

		Estimate	Std. Error	95% I.C	
Fixed					
Intercept		-4.8	3.31	-11.29	1.700
Notification date		0.005	0.001	0.001	0.0079
Risk Group	IDU	0.14	0.01	0.10	0.17
	MSM/Bisexual	-0.03	0.02	-0.065	0.008
Nationality	Americas	-0.06	0.04	-0.14	0.02
	Europe	0.02	0.021	-0.02	0.06
	Low Incidence Countries	0.28	0.16	-0.02	0.59
Stage	AIDS	-0.28	0.012	-0.31	-0.26
Age		-0.003	0.0005	-0.004	-0.002
Recent incidence		0.039	0.0059	0.026	0.054
Random					
Level 3					
(Between Type of Services, 8 levels)	Intercept	std	0.31		
Level 2 (Between Entities, 349 levels)					
	Intercept	std	0.38		
	Recent Incidence	std	0.04		
Level 1					
(Between individuals)		std	0.9		
Number of obs:	26461				

Again, no relevant problems were identified from the profile zeta plot (Figure 5.53).

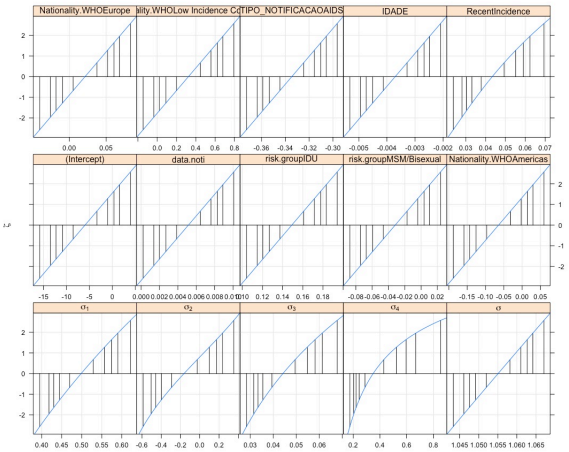


Figure 5.53: Profile zeta plot of the parameters from the model with a random intercept and a random slope.

In Figure 5.54 each line corresponds to a Reporting Institution. In the left panel are presented the estimated intercepts and in the right panel the estimates for the coefficient with respect to Recent Incidence are shown. A large variability is observed by type of services provided (Figure 5.55).

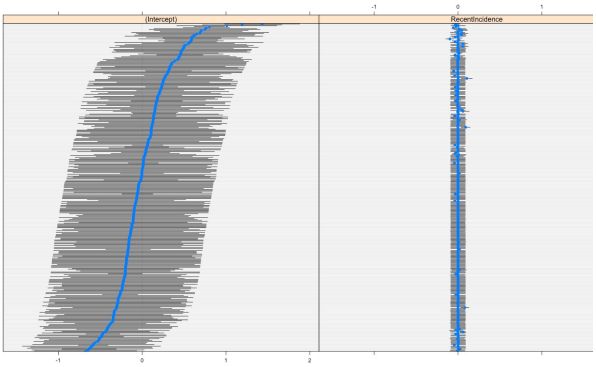


Figure 5.54: 95% prediction intervals on the random effect Entity

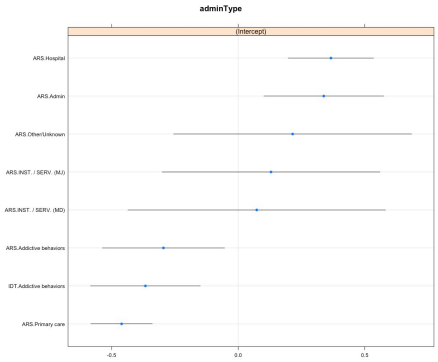


Figure 5.55: 95% prediction intervals on the random effect Administration

In Table 5.16 it is present the comparison between this two models, which revealed that there are significantly different. Moreover, given the AIC, BIC values the model with the random intercept presents as an improvement.

Table 5.16: Multilevel Models comparison

Models:	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
M1: random intercept	13	72073	72179	-36023	72047			
M2: random intercept and random slope	15	70768	70890	-35369	70738	1309.3	2	< 0.001

Concluding, the estimates provided by the null model, the model with a random intercept and the last model with a random intercept and a random slope are presented in Table

5.17. Adding fixed effects to the null model produced an increase on the standard deviations of the random effects - Level 'Type of Services' and ' Entities' and a decrease of the residual standard deviation. When adding the random slope, all standard deviations decreased when comparing to the random intercept model. The comparison of the fixed effect estimates between M1 and M2 models, suggest mild differences between the two models.

Table 5.17: Full model and with random intercept and random intercept and slop

Model		M0: Null	M1: random intercept	M2: random intercept and random slop
Fixed		Coeff.(s.e.)	Coeff.(s.e.)	Coeff.(s.e.)
Intercept		4.67(0.128)	2.714 (3.29)	-4.79 (3.3)
Notification date			0.001 (0.002)	0.005 (0.001)
Risk Group	IDU		0.15(0.015)	0.14 (0.015)
	MSM/Bisexual		-0.04(0.02)	-0.028 (0.019)
Nationality	Americas		-0.074(0.042)	-0.06 (0.04)
	Europe		0.022(0.021)	0.02 (0.02)
	Low Incidence Countries		0.203(0.16)	0.28 (0.16)
Stage	AIDS		-0.31(0.01)	-0.29 (0.013)
Age			-0.0031(0.0006)	-0.003 (0.0006)
Recent incidence			0.002(0.0001)	0.039 (0.006)
Random				
Level 3 (Between Type of Services, 8 levels)		std	std	std
		0.31	0.36	0.31
Level 2 (Between Entities, 349 levels)		std	std	std
Intercept		0.49	0.50	0.38
Recent Incidence				0.04
Level 1 (Between individuals)		std	std	std
		0.98	0.94	0.91
AIC		77085	72073	70768
BIC		77118	72179	70890
logLik		-38538	-36023	-35369
Deviance		77077	72047	70738

5.4 Summary

Since the beginning of the epidemic, the majority of the reported cases were from asymptomatic cases closely followed by the AIDS patients. Following the observed pattern in the European countries, most of the diagnosed cases were men. The heterosexual risk-group was the most common risk-group category. In Portugal, the first case known was diagnosed in 1983 and the first notification was just in 1985. Late diagnosis, that is AIDS, was the most reported stage at the early beginning of the epidemic and of the surveillance system. Around the year 2000 asymptomatic cases were reported more frequently being followed closely by the AIDS cases . The same pattern was observed when this characteristic is represented by the notification year. In the whole epidemic history, males were most frequently observed. IDU ' s were the most frequent diagnosed and reported risk - group until around the year 2002 and then it started to decrease giving place to the heterosexual risk - group. Throughout the years there were several fluctuations on the number of cases per year of notification and RHAs, being Lisbon and the Tagus, and the North the regions with the majority of the cases. It is important to notice a large contribute of cases coming from a non identified source.

Throughout the notification years, the curve of number of notified cases per Region of

Health Administration demonstrated a large variation and with a large number of spikes. The majority of the contributions arose from the RHA LVT and from the RHA North, the two RHA's serving the two major Portuguese cities. From the perspective of the diagnosis year, the epidemic curves were smoother and naturally, the main contributions arose from the same regions.

The observed number of diagnosed HIV-AIDS cases (at all stages) in Portugal, from 1983 to 2011, exhibited an increasing trend between 1983 and the year 2000, four years after the introduction of HAART. Since then, the numbers of diagnosed cases have been steadily decreasing. When the notification became mandatory a slight growth was observed.

Considering the reporting delay classified in two groups $(0, 3]$ months and > 3 months, the descriptive behaviours seemed to be similar with respect to the patient's age at the diagnosis, gender, stage of the disease, transmission risk group, nationality, type of the health care provider that made the diagnosis and administrative and financial responsible from the health care provider.

The reporting delay was firstly studied using the classical approach, developed for large epidemics, more specifically from the CDC data: considering the division into quarters and cross - classifying the quarters of the diagnosis date and considering only AIDS cases. The probability of these cases getting reported is higher than in other stages due to the severe symptoms and the urge of treatment. The model for the reporting delays, based on Poisson counts, estimated that the majority of AIDS cases are reported with 3 months after the diagnosis. And this result was consistent with previous statistical publications using the Portuguese HIV - AIDS data found in [203, 204] and by Mauch when performing the global evaluation of the surveillance system [121]. The assumption of Poisson distribution, although a classical approach, is not unwarranted because HIV - AIDS is a contagious disease and the development of the disease is long and variable and thus recent number of AIDS may be positively correlated with past numbers [69]. Moreover, the model indicated the existence of a slight non-stationary process.

Analysing the longitudinal trajectories of the different quarters of reporting delay (each trajectory is the relative incidence of cases in a particular class of reporting delay), it was found that the trajectory of the quarter $(0, 3]$ months is completely separated from the other trajectories, that can be grouped together. Along the diagnosis year, the trajectories within $(0, 3]$ months always counted with the major percentage of cases. It was observed some fluctuations that ranged from 25% until 75%. Delays longer than 9 months had very close trajectories. These delays had less than 10% of the cases each. So, for modelling, and considering the epidemiological and surveillance relevance, the partitions $(0, 3]$ months and more than 3 months seem to be a good classification for the set of all trajectories.

From the longitudinal model we could conclude that the delay curves have small fluctuations over time. Individuals delay curves can be used to estimate the corresponding delay distribution.

Surveillance systems rely on processes using pre-specified disease case definitions and employ manual data collection, human decision making, and manual data entry. The analysis of incidence and prevalence should thus include the analysis of the historical events that may affect the way systems collect the data.

Considering the reporting delay divided into quarters, a 2-class division arises naturally from the previous analysis, with a cut-off at the 3-month delay. According to this classification, several supervised learning techniques were applied.

Applying data mining in order to build a classification model based on the variable described above, it was applied the Naive Bayes classifier, the K- nearest neighbour algorithm, multilayer perceptrons and support vector machines. The results showed that multilayer perceptrons provided the best results, with a higher classification accuracy (approximately

63%), precision (approximately 76%) and recall (approximately 60%), on the other hand it is considerably slower. It can predict, with reasonable efficiency the group of reporting delays less than 3 months long considering the inputs. The support vector machines model had similar results and was considerable faster.

Since the majority of cases were reported in the reporting delay group $[0, 3]$ months and some fluctuations of the percentage of this pattern were observed, it was decided to model the delays in continuous time and to use a fully parametric approach. This approach avoids the need to, somehow arbitrarily, classify the delays into discrete categories. When the time unit is large, such as a quarter, this can yield very imprecise estimates for the recent incidence. Moreover, by fitting a continuous long tail distribution we are allowing the model to assess the underreporting, since this issue can be considered as a particular case of reporting delay with infinite length. Finally, it was decided to avoid the reverse time proportional hazards regression model in favour of an explicit model for the reporting delay (measured in number of days), using commonly distributions from survival analysis parametric models: normal, log-normal, exponential, Weibull and gamma. This allows a flexible representation of the reporting delay distribution at moderately long delays and accommodates the observed extremely long tails, which are a striking feature of these data [198]. Moreover, if the cases are reported in batches, the multinomial assumption does not capture the pattern and if the discretized time intervals are large this can yield very imprecise estimates for recent HIV-AIDS incidence. Finally, since data is right - truncated (there are no censored reporting delay lengths), the inferences on covariate effects are mainly driven by the shorter delays, which are much more numerous than the longer delays [198]. This is appropriate for rapid detections of changes in epidemic patterns. Within this approach, it was also considered that patients are nested in health care facilities, which are grouped according to its type (hospital, primary care services,...), which in turn are grouped in administrative regions and government by historical policies. After pruning the level of random effects, the final model considered the individuals nested within reporting entities nested into types of services. The final model included fixed-effects from the notification date, the risk group, the nationality, the stage, Age and recent incidence. The distribution of the response was chosen to be the log-normal. From the model and assuming that the values from the remaining covariates are kept constant, it is expected that: any 1-year increase in the notification date increases the reporting delay by 0.5%; any 1-year increase in the patients' age decreases the reporting delay by 0.3% and an 1-patient increase in the absolute incidence increases the reporting delay by 4%. Moreover, an AIDS case is expected to be approximately 25% faster to report than an asymptomatic case. An IDU case is expected to be 14% slower than a Heterosexual while no significant differences regarding the expected reporting delay were identified for MSM/Bisexuals and Heterosexuals.

Finally a note for the BB/NBD model, which is a Bayesian approach with the purpose of capturing the unobserved heterogeneity in reporting AIDS cases. The model was also applied and it was more likely to report an AIDS case when a large number of cases has already been notified. Although the model goodness of fit seems to be reasonable, its estimates do not captured the decreasing pattern that the disease has been showing. One explanation may rely on the fact this is a stop Poisson process without information on the time of diagnosis and treatment effect.

Chapter 6

Conclusions

> In God we trust, all others must bring data.
> **William Edwards Deming (1900-1993)**

This chapter summarizes the contributions and limitations of this thesis, and points out future related work. In order to have timely and accurate representation of the state of the epidemic, the surveillance system must collect data efficiently and effectively. Moreover, it must apply a systematic collection of methods in order to have data with the highest quality possible. This thesis covered some of the essential components of the statistical process flow for modelling and analysing reporting delays for infectious diseases such as HIV-AIDS. This included evaluating and improving data quality, and retrieving knowledge about the reporting delay process in the Portuguese HIV-AIDS surveillance system.

6.1 Synthesis and Main Contributions

The main contributions of this work are concentrated in three distinct areas: concise and detailed description of all the processes leading to the evaluation of infectious diseases incidence numbers; data pre-processing and data quality improvement methods for surveillance data; and definition and implementation of reporting delay modelling methods.

Firstly, the process resulting in the reported incidence numbers was described in a holistic approach. For that purpose, the biology of the virus and its transmission model and the factors influencing HIV - related and / or behaviour - change were described. Focus was given to the processes, which directly or indirectly influence the number of cases observed and reported. These factors can be organized into nested levels: individual, interpersonal / network, institutional / health system and finally structural. The European epidemic was also assessed and the surveillance systems were described in terms of collecting data processes, main stakeholders, and issues rising in each step.

A unified approach for the definition of data quality was described uniting the common issues described in computer science literature and the state-of-the-art of the officials' statistical business processes. A systematic process for performing statistical data pre - processing was proposed, composed by two different components: one for detecting issues and the other for improving data quality with appropriate methods for surveillance data. The methodology was applied to the Portuguese HIV-AIDS Surveillance System.

The classical joint modelling with Poisson counts considering reporting delay and diagnosis date divided into quarters (which were developed in a context of large epidemics) do not fully capture the reporting process of the Portuguese system. These results are consistent with a previous publication using the Portuguese data set but the assumption of Poisson distribution is not unwarranted because HIV - AIDS is a contagious disease and thus recent number of AIDS may be positively correlated with past numbers. Moreover, this

model indicated some non-stationary characteristics of the reporting delay process. When, analysing the reporting delays as a function of the the diagnosis year, two well-separated patterns were identified. In addition, some fluctuations indicated variability in the structure underneath.

Other researchers have also found increases in reporting delay over time that can be attributed largely to changes over time in the distributions of the covariates, or to temporary periods of slower reporting in specific geographical areas that are typically followed by 'catch-up' periods of faster reporting.

For the identification of the main factors influencing the reporting delays in the HIV - AIDS cases within the Portuguese Surveillance System, data - mining methods were applied. The reporting delay was divided into quarters and in a 2-class division with a cut-off at the 3-months delay. According to this classification, several supervised learning techniques were applied and the best performance was obtained by MLP and SVMs. These methods had approximately an accuracy of 60% which is a reasonable result in the biomedical field, and more specifically in the area of reporting delay prediction, and can be explained by the input data characteristics.

Since most of the cases were reported in the reporting delay $[0, 3]$ months and some fluctuations of the percentage of this pattern were observed, it was decided to model delays in continuous time and to use a fully parametric approach allowing a more flexible approach. The right - truncated characteristic was accommodated naturally in the sense that it was allowed that the inferences on covariate effects were mainly driven by the shorter delays which are appropriate when the goal is to rapidly detect changes in epidemic patterns. The model was improved by adding the context considering the patient nationality, age, risk group, the notification date and recent incidence observed in each institution. It was taken into consideration that patients are nested within health care facilities, which in turn are grouped according to their type (hospital, primary care services,...), which are grouped in administrative regions and governed by historical policies. A maximum model with different theoretical distributions for the (conditional) response and different nested contexts was fitted and, afterwards, this complex model was pruned.

The under-report issue was addressed in two different ways: with a Bayesian approach and with no covariate effects and integrated with the continuous multilevel model as a particular case of reporting delay with infinite length.

All the multilevel models presented in the thesis were also fitted using Bayesian approaches, with very similar results to the ones here exposed. A simple probabilistic record linkage between the Portuguese HIV-AIDS data set and with the Portuguese tuberculosis data set was also performed but the data available on both data sets did not enable an in depth analysis.

6.2 Limitations

The main limitation of this work is related with the data specific characteristics and the data collection method. It is widely recognized that the quality of the data within the biomedical and healthcare fields is inferior to that found in other fields. In our data set the main reasons for the poor classification quality are most probably related to stigma around the disease that leads patients to provide incorrect information (mainly in the transmission group), to high demands of the healthcare systems and to implementation of the surveillance system, more specifically paper form reports and poor communication between the stakeholders. In a previous qualitative assessment of the Portuguese Surveillance System, Mauch in [93] pointed out that all clinicians reported that they complete the notification form after the patient has left the office, sometimes several days or weeks later. This practice has the potential to contribute to inaccuracies in reporting for some variables, such as the associated

risk group, due to recall errors. Moreover, errors may also arise from the transcription of the information from the paper report to computer databases.

6.3 Recommendations for Future Work

Inspired by the work made in this thesis, a number of interesting fields of investigation were open. In the following paragraphs, several recommendations for future work are suggested:

- Creation of a longitudinal missing data imputation method that can capture the data missing mechanism and accommodate changes in the surveillance system.
- Extension of the BB/NBD model in order to capture the data heterogeneity and that may work properly, in a more stable and reliable manner.
- Comparison of the methodologies expressed in this thesis with others and with the ones defined for epidemics of other countries, taking into consideration that the data collection methods and culture are the main differentiating keys.
- Fit the multilevel models with other distributions such as Tweedie distributions which have a higher point mass at zero.
- Extend the multilevel model by adding a degree of (un)confidence to each variable, independently of the type in which it they are measured.
- Extend the approach allowing ensembling methods.
- Apply the methods developed for data pre-processing and report delay modelling for other infectious diseases surveillance data such as Tuberculosis.

All these perspectives for future work could lead to a better understanding and modelling of the data. However, the statistical data preprocessing and reporting delay estimation methods, proposed in this thesis are a concise first step, in the improvement of infectious diseases surveillance systems, towards the creation of more effective epidemics control procedures.

Bibliography

[1] E. Choffnes, P. Sparling, M. Hamburg, S. Lemon, A. Mack, et al. *Global Infectious Disease Surveillance and Detection:: Assessing the Challenges–Finding Solutions, Workshop Summary*. National Academies Press, 2007.

[2] World Health Organization and others. Communicable disease surveillance and response systems: guide to monitoring and evaluating. Technical report, World Health Organization, 2006.

[3] European Centre for Disease Prevention and Control. Data quality monitoring and surveillance system evaluation. Technical report, ECDC, 2014.

[4] M. Stoto. Syndromic surveillance in public health practice. In *Institute of Medicine, ed. Infectious Disease Surveillance and Detection (Workshop Report)*, pages 63–72, 2007.

[5] R. Jajosky and S. Groseclose. Evaluation of reporting timeliness of public health surveillance systems for infectious diseases. *BMC public health*, 4(1):29, 2004.

[6] T. Doyle, M. Glynn, and S. Groseclose. Completeness of notifiable infectious disease reporting in the united states: an analytical literature review. *American journal of epidemiology*, 155(9):866–874, 2002.

[7] I. Medicine, B. Prevention, P. Evaluation, and C. Allocation. *Measuring What Matters: Allocation, Planning, and Quality Assessment for the Ryan White CARE Act*. National Academies Press, 2004.

[8] M. Pagano, X. Tu, V. De Gruttola, and S. MaWhinney. Regression analysis of censored and truncated data: estimating reporting-delay distributions and AIDS incidence from surveillance data. *Biometrics*, pages 1203–1214, 1994.

[9] World Health Organization et al. WHO report on global surveillance of epidemic-prone infectious diseases, 2000.

[10] A. Fauci. Infectious diseases: considerations for the 21st century. *Clinical Infectious Diseases*, 32(5):675–685, 2001.

[11] A. Fauci, N. Touchette, and G. Folkers. Emerging infectious diseases: a 10-year perspective from the national institute of allergy and infectious diseases. *International Journal of Risk & Safety in Medicine*, 17(3, 4):157–167, 2005.

[12] S. Teutsch and R. Churchill. *Principles and Practice of Public Health Surveillance*. Oxford University Press, 2000.

[13] M. Stoto. Public Health Surveillance: A historical review with a focus on HIV/AIDS. *Santa Monica, Calif, USA*, 2003.

[14] L. Waller and C. Gotway. *Applied spatial statistics for public health data*, volume 368. John Wiley & Sons, 2004.

[15] H. Chen, D. Hailey, N. Wang, and P. Yu. A review of data quality assessment methods for public health information systems. *International Journal of Environmental Research and Public Health*, 11(5):5170–5207, May 2014.

[16] A. Karr, A. Sanil, and D. Banks. Data quality: A statistical perspective. *Statistical Methodology*, 3(2):137–173, 2006.

[17] P. McCullagh and J. Nelder. *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1989.

[18] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[19] J. Gonçalves, B. Faria, L. Reis, V. Carvalho, and A. Rocha. Data mining and electronic devices applied to quality of life related to health data. In *Information Systems and Technologies (CISTI), 2015 10th Iberian Conference on*, pages 1–4. IEEE, 2015.

[20] A. Oliveira, P. Machado, C. Rebelo, R. Gaio, and L. Reis. Data quality mining in health systems. 2018.

[21] A. Oliveira, P. Machado, C. Rebelo, R. Gaio, and L. Reis. Statistical pre - processing in surveillance systems - an application to portuguese HIV - AIDS surveillance data. 2018.

[22] A. Oliveira, J. Costa, and R. Gaio. The incidence of AIDS in Portugal adjusted for reporting delay and underreporting. In *Information Systems and Technologies (CISTI), 2014 9th Iberian Conference on*, pages 1–5. IEEE, 2014.

[23] A. Oliveira, R. Gaio, J. da Costa, and L. Reis. An approach for assessing the distribution of reporting delay in portuguese AIDS data. In *New Advances in Information Systems and Technologies*, pages 641–649. Springer, 2016.

[24] A. Oliveira, B. Faria, R. Gaio, and L. Reis. Data mining in HIV-AIDS surveillance system. *Journal of Medical Systems*, 41(4):51, 2017.

[25] A. Oliveira, H. Amorim, R. Gaio, and L. Reis. Multilevel modelling of HIV - AIDS reporting. 2018.

[26] A. Sousa, P.and Oliveira, R. Gaio, and R. Duarte. Longitudinal clustering of tuberculosis incidence in Portugal from 2002 to 2012. *European Respiratory Journal*, 46(suppl 59):OA1450, 2015.

[27] A. Oliveira, A. Salgado, A. Magalhães, J. Faias, and L. Miranda. Perceived quality of life among first-year health students engaged in problem based learning and traditional teaching model: First-year students of allied health sciences. In *Information Systems and Technologies (CISTI), 2016 11th Iberian Conference on*, pages 1–7. IEEE, 2016.

[28] S. Dübel and J. Reichert. *Handbook of therapeutic antibodies*, volume 1. John Wiley & Sons, 2014.

[29] A. Signore. About inflammation and infection. *EJNMMI research*, 3(1):1–2, 2013.

[30] A. Abbas, A. Lichtman, and S. Pillai. *Basic immunology: functions and disorders of the immune system*. Elsevier Health Sciences, 2012.

[31] T. Kindt, Ri. Goldsby, B. Osborne, and J. Kuby. *Kuby immunology*. Macmillan, 2007.

[32] P. Shetty. *Nutrition, Immunity and Infection*. Modular texts. CABI, 2010.

[33] World Health Organization. The top 10 causes of death. fact sheet n 310, May 2014.

[34] A. Fauci. Pathogenesis of HIV disease: opportunities for new prevention interventions. *Clinical Infectious Diseases*, 45(Supplement 4):S206–S212, 2007.

[35] J. Hutchinson. The biology and evolution of HIV. *Annual review of anthropology*, pages 85–108, 2001.

[36] A. Harries, D. Maher, and S. Graham. *TB/HIV: a clinical manual*. World Health Organization, 2004.

[37] World Health Organization and others. *WHO case definitions of HIV for surveillance and revised clinical staging and immunological classification of HIV-related disease in adults and children*. Geneva: World Health Organization, 2007.

[38] A. Fauci, G. Pantaleo, S. Stanley, and D. Weissman. Immunopathogenic mechanisms of HIV Infection. *Annals of Internal Medicine*, 124(7):654–663, 1996.

[39] World Health Organization. HIV/AIDS. fact sheet n 360, 2014.

[40] K. Jacobsen. *Introduction to Global Health*. Jones & Bartlett Learning, 2009.

[41] J. Tang, X. Li, M. Price, E. Sanders, O. Anzala, E. Karita, A. Kamali, S. Lakhi, S. Allen, E. Hunter, et al. CD4: CD8 lymphocyte ratio as a quantitative measure of immunologic health in HIV-1 infection: findings from an African cohort with prospective data. *Frontiers in microbiology*, 6:670, 2015.

[42] Centers for Disease Control and Prevention. HIV transmission, December 2015.

[43] Joint United Nations Programme on HIV/AIDS and others. A framework for monitoring and evaluating HIV prevention programmes for most-at-risk populations, 2008.

[44] Joint United Nations Programme on HIV/AIDS and World Health Organization. Guidelines on surveillance among populations most-at-risk for HIV, 2011.

[45] S. Deeks, S. Lewin, and D. Havlir. The end of AIDS: HIV infection as a chronic disease. *The Lancet*, 382(9903):1525–1533, 2013.

[46] World Health Organization. Guideline on when to start antiretroviral therapy and on pre-exposure prophylaxis for HIV. Technical report, World Health Organization, september 2015.

[47] The Lancet Infectious Diseases. Co-infection: new battlegrounds in HIV/AIDS. *The Lancet Infectious Diseases*, 13(7):559, 2013.

[48] World Health Organization. A deadly partnership: tuberculosis in the era of HIV. consensus statement. geneva. Technical Report WHO/TB/96.204, World Health Organization, 2002.

[49] T. Shinnick. *Tuberculosis*. Current Topics in Microbiology and Immunology. Springer Berlin Heidelberg, 2013.

[50] A. Cooper. Cell-mediated immune responses in tuberculosis. *Annual review of immunology*, 27:393–422, 2009.

[51] World Health Organization and others. Global tuberculosis report 2015. Technical report, World Health Organization, 2015.

[52] R. Koch. The etiology of tuberculosis. *Review of Infectious Diseases*, 4(6):1270–1274, 1982.

[53] Centers for Disease Control and Prevention. TB elimination - tuberculosis: General information, October 2011.

[54] C. Dyer. *Tuberculosis*. Biographies of disease. Greenwood, 2010.

[55] World Health Organization. *Global Tuberculosis Report 2015*. Global tuberculosis control. World Health Organization, 2016.

[56] A. White, L. Sibley, M. Dennis, K. Gooch, G. Betts, N. Edwards, A. Reyes-Sandoval, M. Carroll, A. Williams, P. Marsh, et al. Evaluation of the safety and immunogenicity of a candidate tuberculosis vaccine, MVA85A, delivered by aerosol to the lungs of macaques. *Clinical and Vaccine Immunology*, 20(5):663–672, 2013.

[57] J. Mitra, K. Vinayak, and R. Mitra. Study of clinical profile of tuberculosis in HIV infected patients with special reference to CD4 T cell count and its oral manifestation. *International Journal of Medical Research and Review*, 4(06), 2016.

[58] A. Pawlowski, M. Jansson, M. Skold, M. Rottenberg, and G. Kallenius. Tuberculosis and HIV Co-infection. *PLoS Pathog*, 8(2):1–7, 02 2012.

[59] E. Corbett, C. Watt, N. Walker, D. Maher, B. Williams, M. Raviglione, and C. Dye. The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. *Archives of internal medicine*, 163(9):1009–1021, 2003.

[60] World Health Organization and others. A guide to monitoring and evaluation for collaborative TB/HIV activities - 2015 revision. Technical report, World Health Organization, 2015.

[61] J. Lazarus, M. Olsen, L. Ditiu, and S. Matic. Tuberculosis - HIV co-infection: policy and epidemiology in 25 countries in the WHO European region. *HIV medicine*, 9(6):406–414, 2008.

[62] M. van der Werf, C. Ködmön, P. Zucs, V. Hollo, A. Amato-Gauci, and A. Pharris. Tuberculosis and HIV coinfection in Europe: looking at one reality from two angles. *AIDS (London, England)*, 30(18):2845, 2016.

[63] Doentes com VIH vão fazer tratamentos de prevenção o para a tuberculose, March 2015.

[64] World Health Organization. Health in 2015: from MDGs to SDGs. Technical report, World Health Organization, 2015.

[65] World Health Organization and others. Progress report 2016: prevent HIV, test and treat all: WHO support for country impact, 2016.

[66] European Centre for Disease Prevention and Control/WHO Regional Office for Europe. HIV/AIDS surveillance in europe 2015. Technical report, Stockholm: ECDC, 2016.

[67] European Centre for Disease Prevention and Control/WHO Regional Office for Europe. HIV/AIDS surveillance in europe 2014. Technical report, Stockholm: ECDC, 2015.

[68] L. Platt, E. Jolley, V. Hope, A. Latypov, P. Vickerman, F. Hickson, L. Reynolds, and T. Rhodes. *HIV epidemics in the European region: vulnerability and response*. Directions in Development. International Bank for Reconstruction and Development/The World Bank, 2015.

[69] European Centre for Disease Prevention and Control/WHO Regional Office for Europe. HIV/AIDS surveillance in europe 2011. Technical report, Stockholm: ECDC, 2012.

[70] M. Kaufman, F. Cornish, R. Zimmerman, and B. Johnson. Health behavior change models for HIV prevention and AIDS care: practical recommendations for a multilevel approach. *Journal of acquired immune deficiency syndromes (1999)*, 66(Suppl 3):S250, 2014.

[71] J. Parkhurst. HIV prevention, structural change and social values: the need for an explicit normative approach. *Journal of the International AIDS Society*, 15(3), 2012.

[72] S. Baral, C. Logie, A. Grosso, A. Wirtz, and C. Beyrer. Modified social ecological model: a tool to guide the assessment of the risks and risk contexts of HIV epidemics. *BMC public health*, 13(1):1, 2013.

[73] S. Kippax. Understanding and integrating the structural and biomedical determinants of HIV infection: a way forward for prevention. *Current Opinion in HIV and AIDS*, 3(4):489–494, 2008.

[74] E. Sumartojo. Structural factors in HIV prevention: concepts, examples, and implications for research. *Aids*, 14:S3–S10, 2000.

[75] Centers for Disease Control and Prevention and Centers for Disease Control and Prevention and others. The social-ecological model: A framework for prevention. *Injury center: Violence prevention*, 2009.

[76] R. Valdiserri. HIV/AIDS stigma: an impediment to public health. *American journal of public health*, 92(3):341–342, 2002.

[77] L. Collins, K. Kugler, and M. Gwadz. Optimization of multicomponent behavioral and biobehavioral interventions for the prevention and treatment of HIV/AIDs. *AIDS and Behavior*, 20(1):197–214, 2016.

[78] A. Desclaux. Stigmatization and discrimination: What does a cultural approach have to offer. *HIV/AIDS, stigma and discrimination: an anthropological approach*, pages 1–10, 2003.

[79] S.a Gari, C. Doig-Acuña, T. Smail, J. Malungo, A. Martin-Hilber, and S. Merten. Access to HIV/AIDS care: a systematic review of socio-cultural determinants in low and high income countries. *BMC Health Services Research*, 13(1):198, 2013.

[80] A. Mahajan, J. Sayles, V. Patel, R. Remien, D. Ortiz, G. Szekeres, and T. Coates. Stigma in the HIV/AIDS epidemic: a review of the literature and recommendations for the way forward. *AIDS (London, England)*, 22(Suppl 2):S67, 2008.

[81] European Centre for Disease Prevention and Control. Evidence brief: Impact of stigma and discrimination on access to HIV services in europe. monitoring implementation of the dublin declaration on partnership to fight HIV/AIDS in europe and central asia. Technical report, Stockholm: ECDC, 2017.

[82] M. Mondal and M. Shitan. Factors affecting the HIV/AIDS epidemic: an ecological analysis of global data. *African health sciences*, 13(2):301–310, 2013.

[83] B. Yakob and B. Ncama. A socio-ecological perspective of access to and acceptability of HIV/AIDS treatment and care services: a qualitative case study research. *BMC Public Health*, 16(1):155, 2016.

[84] European Centre for Disease Prevention and Control. Evidence brief: HIV and laws and policies in europe. monitoring implementation of the dublin declaration on partnership to fight HIV/AIDS in europe and central asia. Technical report, Stockholm: ECDC, 2017.

[85] Centers for Disease Control and Prevention and others. Principles of epidemiology in public health practice: an introduction to applied epidemiology and biostatistics, 2006.

[86] European Centre for Disease Prevention and Control. Policy on data submission, access, and use of data within tessy – 2015 revision. Technical report, Stockholm: ECDC, 2015.

[87] European Centre for Disease Prevention and Control. Indicator-based surveillance, 2005-2016.

[88] Centre for Disease Prevention and Control. NMI Overview, 2016.

[89] P. MacDonald. *Methods in Field Epidemiology*. Jones & Bartlett Learning, 2011.

[90] C. Gibbons, M. Mangen, D. Plass, A. Havelaar, R. Brooke, P. Kramarz, K. Peterson, A. Stuurman, A. Cassini, E. Fèvre, and M. Kretzschmar. Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. *BMC Public Health*, 14(1):1–17, 2014.

[91] A. Marinovic, C. Swaan, J. van Steenbergen, and M. Kretzschmar. Quantifying reporting timeliness to improve outbreak control. *Emerging Infectious Diseases*, 21(2):209–216, February 2015.

[92] E. Reijn, C. Swaan, M. Kretzschmar, and J. van Steenbergen. Analysis of timeliness of infectious disease reporting in the netherlands. *BMC Public Health*, 11(1):409, 2011.

[93] S. Mauch. Situational assessment of the HIV/AIDS notification system - a Portuguese experience. Technical report, National Coordination For HIV Infection, July 2009.

[94] R. Gliklich, N. Dreyer, and M. Leavy. Registries for evaluating patient outcomes: A user's guide [internet]. 3rd edition. 2014 apr. 11, data collection and quality assurance, 2014.

[95] P. Anthamatten and H. Hazen. *An introduction to the geography of health*. Routledge, 2012.

[96] W. Trick. Decision making during healthcare-associated infection surveillance: A rationale for automation. *Clinical Infectious Diseases*, 57(3):434, 2013.

[97] N. Walker, J. Garcia-Calleja, L. Heaton, E. Asamoah-Odei, G. Poumerol, S. Lazzari, P. Ghys, B. Schwartländer, and K. Stanecki. Epidemiological analysis of the quality of HIV sero-surveillance in the world: how well do we track the epidemic? *AIDS*, 15(12):1545–1554, 2001.

[98] European Commission. Commission implementing decision of 8 august 2012 amending decision 2002/253/ec laying down case definitions for reporting communicable diseases to the community network under decision no 2119/98/ec of the european parliament and of the council. *Official Journal of the European Union*, 55, 2012.

[99] U.S. Department of Health and Human Services. *Guide for HIV/AIDS Clinical Care*. Health Resources and Service Administration, Rockville, MD: U.S. Department of Health and Human Services, 2014 edition edition, 2014.

[100] K. Castro, J. Ward, L. Slutsker, J. Buehler, H. Jaffe, R. Berkelman, and J. Curran. 1993 revised classification system for HIV infection and expanded surveillance case definition for AIDS among adolescents and adults. *Clinical Infectious Diseases*, 17(4):802–810, 1993.

[101] I. Devaux, J. Alix, G. Likatavicius, M. Herida, S. Nielsen, F. Hamers, and A. Nardone. Human immunodeficiency virus and acquired immunodeficiency syndrome case reporting in the World Health Organization European Region in 2006. *Euro surveillance: bulletin Europeen sur les maladies transmissibles= European communicable disease bulletin*, 13(39):932–946, 2008.

[102] EuroHIV. Report on the EuroHIV 2006 survey on HIV and AIDS surveillance in the WHO European Region. Technical report, Saint-Maurice: Institut de veille sanitaire, 2007.

[103] M. Rosinska, N. Pantazis, J. Janiec, A. Pharris, A. Amato-Gauci, C. Quinten, and ECDC HIV/AIDS Surveillance Network. Missing data and reporting delay in the European HIV Surveillance System: exploration of potential adjustment methodology. 2017.

[104] European Centre for Disease Prevention and Control. HIV testing in europe. evaluation of the impact of the ECDC guidance on HIV testing: increasing uptake and effectiveness in the european union. Technical report, Stockholm: ECDC, 2016.

[105] European Centre for Disease Prevention and Control. HIV/AIDS reporting protocol and analysis plan 2016- surveillance data for 2015. Technical report, European Centre for Disease Prevention and Control and World Health Organization - Regional Office for Europe, 2016.

[106] European Centre for Disease Prevention and Control. Annual epidemiological report 2014 - sexually transmitted infections, including HIV and blood-borne viruses. Technical report, Stockholm: ECDC, 2015.

[107] Ministerio da Saude. Portaria n. 258/2005 de 16 de marco. *DIÁRIO DA REPÚBLICA — I SÉRIE-B*, 2005.

[108] T. Shivaji and H. Martins. Improving data management practices in the portuguese HIV/AIDS surveillance system during a time of public sector austerity. *BMJ Quality Improvement Reports*, 4(1):u209037.w3663, August 2015.

[109] P. Barros, S. Machado, and J. Simões. Portugal: health system review. *Health systems in transition*, 13(4):1–156, 2011.

[110] Ministério da Saúde. A organização interna e a governação dos hospitais, 2010.

[111] A. Paulo. SNS: Caracterização e desafios. *GPEARI-MFAP. Lisboa, Setembro, 16pp*, 2010.

[112] M. Oliveira and G. Bevan. Measuring geographic inequities in the Portuguese health care system: an estimation of hospital care needs. *Health policy*, 66(3):277–293, 2003.

[113] M. Oliveira and C. Pinto. Health care reform in Portugal: an evaluation of the NHS experience. *Health Economics*, 14(S1), 2005.

[114] C. Encarna¸cao, F. Amado, and s. Santos. Challenges for performance assessment and improvement in primary health care: The case of the portuguese health centres. *Health Policy*, 91(1):43 – 56, 2009.

[115] J. Chinen and W. Shearer. Secondary immunodeficiencies, including HIV infection. *Journal of Allergy and Clinical Immunology*, 125(2):S195–S203, 2010.

[116] J. Barnard and X. Meng. Applications of multiple imputation in medical studies: from AIDS to NHANEs. *Statistical Methods in Medical Research*, 8(1):17–36, 1999.

[117] R. Brookmeyer and M. Gail. *AIDS Epidemiology: A Quantitative Approach*. Monographs in Epidemiology and Biostatistics. Oxford University Press, 1994.

[118] P. Bouman, V. Dukic, and X. Meng. A bayesian multiresolution hazard model with application to an AIDS reporting delay study. *Statistica Sinica*, pages 325–357, 2005.

[119] J. Cui and J. Kaldor. Changing pattern of delays in reporting AIDS diagnoses in australia. *Australian and New Zealand journal of public health*, 22(4):432–435, 1998.

[120] T. Green. Using surveillance data to monitor trends in the AIDS epidemic. *Statistics in Medicine*, 17(2):143–154, 1998.

[121] J. Harris. Reporting delays and the incidence of AIDS. *Journal of the American Statistical Association*, 85(412):915–924, 1990.

[122] J. Kalbfleisch and J. Lawless. Regression models for right truncated data with applications to AIDS incubation times and reporting lags. *Statistica Sinica*, 1:19–32, 1991.

[123] B. Balasingam, P. Mannaru, D. Sidoti, K. Pattipati, and P. Willett. *Online Anomaly Detection in Big Data: The First Line of Defense Against Intruders*, pages 83–107. Springer International Publishing, Cham, 2017.

[124] L. Cai and Y. Zhu. The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 14, 2015.

[125] H. Wickham. Tidy data. *Journal of Statistical Software*, 59(10):1–23, 2014.

[126] T. Dasu and T. Johnson. *Exploratory data mining and data cleaning*, volume 479. John Wiley & Sons, 2003.

[127] S. Keller, G. Korkmaz, M. Orr, A. Schroeder, and S. Shipp. The evolution of data quality: Understanding the transdisciplinary origins of data quality concepts and approaches. *Annual Review of Statistics and Its Application*, 4:85–108, 2017.

[128] N. Laranjeiro, S. Soydemir, and J. Bernardino. A survey on data quality: classifying poor data. In *Dependable Computing (PRDC), 2015 IEEE 21st Pacific Rim International Symposium on*, pages 179–188. IEEE, 2015.

[129] BARC (Business Application research Center) - a CXP Group Company. Data quality and master data management: How to improve your data quality, 2017.

[130] EUROSTAT. *Handbook on Data Validation in Eurostat -Practical Guide to Data Validation in EuroSttat*. 2010.

[131] M. Azimaee, M. Smith, L. Lix, C. Burchill, and J. Orr. Mchp data quality framework. *Winnipeg (Manitoba): Manitoba Centre for Health Policy, University of Manitoba*, 2015.

[132] E. De Jonge and M. van der Loo. An introduction to data cleaning with r. *Heerlen: Statistics Netherlands*, 2013.

[133] E. Rahm and H. Do. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4):3–13, 2000.

[134] J. Hipp, U. Güntzer, and U. Grimmer. Data quality mining-making a virute of necessity. In *DMKD*, 2001.

[135] I. Taleb, R. Dssouli, and M. Serhani. Big data pre-processing: A quality framework. In *Big Data (BigData Congress), 2015 IEEE International Congress on*, pages 191–198. IEEE, 2015.

[136] ESS Task Force Peer Review. *Quality Assurance Framework of the European Statistical System- Version 1.2*, 2015.

[137] J. Barateiro and H. Galhardas. A survey of data quality tools. *Datenbank-Spektrum*, 14(15-21):48, 2005.

[138] M. van der Loo. A formal typology of data validation functions. 2015.

[139] A. Chalamalla, I. Ilyas, M. Ouzzani, and P. Papotti. Descriptive and prescriptive data cleaning. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 445–456. ACM, 2014.

[140] G. Cong, W. Fan, F. Geerts, X. Jia, and S. Ma. Improving data quality: Consistency and accuracy. In *Proceedings of the 33rd international conference on Very large data bases*, pages 315–326. VLDB Endowment, 2007.

[141] M. Zio, N. Fursova, T. Gelsema, S. Giebing, U. Guarnera, J. Petrauskiene, Q. Kalben, M. Scanu, K. Bosch, M. van der Loo, and K. Walsdorfer. *Methodology for data validation 1.0*, 2016.

[142] M. van der loo. Properties of validation rules. In *Methodology for data validation 1.0*. 2016.

[143] M. van der Loo. Validation levels based on decomposition of metadata - Essnet Validat Foudation. In *Methodology for data validation 1.0*. 2016.

[144] M. van der Loo. Validation levels from a business prespective - essnet validat foudation. In *Methodology for data validation 1.0*. 2016.

[145] E. de Jonge and M. van der Loo. Error localization as a mixed integer problem with the editrules package. Technical report, Technical Report 2012XX, Statistics Netherlands, The Hague. forthcoming, 2014.

[146] W. Winkler. Methods for evaluating and creating data quality. *Information Systems*, 29(7):531–550, 2004.

[147] A. de Waal. *Processing of erroneous and unsafe data*. Number ERIM PhD Series; EPS-2003-024-LIS. 2003.

[148] T. De Waal, J. Pannekoek, and S. Scholtus. *Handbook of statistical data editing and imputation*, volume 563. John Wiley & Sons, 2011.

[149] J. Osborne and A. Overbay. The power of outliers (and why researchers should always check for them). *Practical assessment, research & evaluation*, 9(6):1–12, 2004.

[150] J. Schafer and J. Graham. Missing data: our view of the state of the art. *Psychological methods*, 7(2):147, 2002.

[151] F. Harrell. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Graduate Texts in Mathematics. Springer, 2001.

[152] S. Dusetzina, S. Tyree, A. Meyer, A. Meyer, L. Green, and W. Carpenter. *Linking data for health services research: a framework and instructional guide*. Agency for Healthcare Research and Quality (US), Rockville (MD), 2014.

[153] B. Forchhammerl, T. Papenbrockl, T. Steningl, and S. Viehmeierl. Duplicate detection on GPUs. *HPI Future SOC Lab: proceedings 2011*, 70:59, 2013.

[154] A. Elmagarmid, P. Ipeirotis, and V. Verykios. Duplicate record detection: A survey. *Knowledge and Data Engineering, IEEE Transactions on*, 19(1):1–16, 2007.

[155] M. van der Loo. The stringdist package for approximate string matching, 2014.

[156] P. Christen. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media, 2012.

[157] S. Serrano-Villar, S. Moreno, M. Fuentes-Ferrer, C. Sánchez-Marcos, M. Avila, T. Sainz, N. Villar, A. Fernández-Cruz, and V. Estrada. The CD4: CD8 ratio is associated with markers of age-associated disease in virally suppressed HIV-infected patients with immunological recovery. *HIV medicine*, 15(1):40–49, 2014.

[158] E. de Jonge and M. van der Loo. *errorlocate: Locate Errors with Validation Rules*, 2016. R package version 0.1.2.

[159] M. van der Loo and E. de Jonge. *Statistical Data Cleaning with Applications in R*. John Wiley & Sons, 2018.

[160] W. Winkler. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage, 1990.

[161] M. Hermans and F. Schadd. A generalization of the winkler extension and its application for ontology mapping. In *BNAIC 2012 The 24th Benelux Conference on Artificial Intelligence*, page 123, 2012.

[162] R. Magnani, K. Sabin, T. Saidel, and D. Heckathorn. Review of sampling hard-to-reach and hidden populations for HIV surveillance. *Aids*, 19:S67–S72, 2005.

[163] Y. Xia, N. Lu, I. Katz, R. Bossarte, J. Arora, H. He, J. Tu, B. Stephens, A. Watts, and X. Tu. Models for surveillance data under reporting delay: applications to US veteran first-time suicide attempters. *Journal of applied statistics*, 42(9):1861–1876, 2015.

[164] A. Agresti. *An Introduction to Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley, 2007.

[165] A. Dobson. *An introduction to generalized linear models*. CRC press, 2001.

[166] J. De Leeuw, E. Meijer, and H. Goldstein. *Handbook of multilevel analysis*. Springer, 2008.

[167] M. Turkman and G. Silva. Modelos lineares generalizados - da teoria à prática. In *VIII Congresso Anual da Sociedade Portuguesa de Estatística, Lisboa*, 2000.

[168] J. Faraway. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*, volume 66. Chapman & Hall/CRC, 2005.

[169] A.w Gelman and J. Hill. *Data analysis using regression and multilevelhierarchical models*, volume 1. Cambridge University Press New York, NY, USA, 2007.

[170] B. Tabachnick and L. Fidell. *Using multivariate statistics*. Allyn & Bacon/Pearson Education, 2007.

[171] P. Wang, Y. Li, and C. Reddy. Machine learning for survival analysis: A survey. *arXiv preprint arXiv:1708.04649*, 2017.

[172] S.a Jacob and R. Ramani. Data mining in clinical data sets: a review. *training*, 4(6), 2012.

[173] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J. Chang, and L. Hua. Data mining in healthcare and biomedicine: a survey of the literature. *Journal of medical systems*, 36(4):2431–2448, 2012.

[174] M. Dunham. *Data Mining: Introductory And Advanced Topics*. Pearson Education, 2006.

[175] J. Iavindrasana, G. Cohen, A. Depeursinge, H. Müller, R. Meyer, and A. Geissbuhler. Clinical data mining: a review. *Yearbook of medical informatics*, 18(01):121–133, 2009.

[176] J. Lindsey. *Applying generalized linear models*. Springer, 1997.

[177] B. George, S. Seals, and I. Aban. Survival analysis and regression models. *Journal of Nuclear Cardiology*, 21(4):686–694, 2014.

[178] J. Pinheiro and D. Bates. *Mixed-Effects Models in S and S-PLUS*. Statistics and Computing. Springer New York, 2006.

[179] D. Bates. lme4: Mixed-effects modeling with r, 2010.

[180] Y. Dodge. *The Oxford dictionary of statistical terms*. Oxford University Press on Demand, 2006.

[181] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[182] C. Genolini and B. Falissard. Kml: k-means for longitudinal data. *Computational Statistics*, 25(2):317–328, 2010.

[183] M. Murty and V. Devi. *Pattern recognition: An algorithmic approach*. Springer Science & Business Media, 2011.

[184] G. John and P. Langley. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pages 338–345, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[185] R. Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[186] T. Sharaf and C. Tsokos. Two artificial neural networks for modeling discrete survival time of censored data. *Adv. in Artif. Intell.*, 2015:1:1–1:1, January 2015.

[187] Y. Yang. Neural network survival analysis. Master's thesis, Faculty of Sciences - Department of Applied mathematics and computer science, 2011.

[188] M. Paliwal and U. Kumar. Assessing the contribution of variables in feed forward neural network. *Applied Soft Computing*, 11(4):3690–3696, 2011.

[189] G. Yin and Y. Ma. Pearson-type goodness-of-fit test with bootstrap maximum likelihood estimation. *Electron. J. Statist.*, 7:412–427, 2013.

[190] C. Schunn and D. Wallach. Evaluating goodness-of-fit in comparison of models to data. *Psychologie der Kognition: Reden and Vorträge anlässlich der Emeritierung von Werner Tack*, pages 115–154, 2005.

[191] J. Hox, M. Moerbeek, and R. van de Schoot. *Multilevel Analysis: Techniques and Applications, Second Edition*. Quantitative Methodology Series. Taylor & Francis, 2010.

[192] C. Judd, G. McClelland, and C. Ryan. *Data analysis: a model comparison approach to regression, ANOVA, and beyond*. Routledge, 2017.

[193] B. Sribney. Problems with stepwise regression, 1996.

[194] D. Altman and P. Andersen. Bootstrap investigation of the stability of a cox regression model. *Statistics in medicine*, 8(7):771–783, 1989.

[195] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[196] S. Derksen and H. Keselman. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, 45(2):265–282, 1992.

[197] D. Faraggi and R. Simon. A neural network model for survival data. *Statistics in medicine*, 14(1):73–82, 1995.

[198] A. Noufaily, Y. Ghebremichael-Weldeselassie, D. Enki, P. Garthwaite, N. Andrews, A. Charlett, and P. Farrington. Modelling reporting delays for outbreak detection in infectious disease data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(1):205–222, 2015.

[199] R. Brookmeyer and J. Liao. The analysis of delays in disease reporting: methods and results for the acquired immunodeficiency syndrome. *American Journal of Epidemiology*, 132(2):355–365, 1990.

[200] J. Kalbfleisch. *Probability and Statistical Inference: Probability. vol. 1*. Probability and statistical inference. Springer-Verlag, 1985.

[201] S. Lagaskos, L. Barraj, and V. Gruttola. Nonparametric analysis of truncated survival data, with application to AIDS. *Biometrika*, 75(3):515–523, 1988.

[202] R. Brookmeyer and J. Liao. Statistical modelling of the AIDS epidemic for forecasting health care needs. *Biometrics*, pages 1151–1163, 1990.

[203] J. Amaral, M. Rosário, and M. Paixao. Data and projections of HIV and AIDS in portugal. *Journal of Applied Statistics*, 27(3):269–279, 2000.

[204] J. Amaral, E. Pereira, and M. Paixão. Data and projections of HIV/AIDS cases in portugal: an unstoppable epidemic? *Journal of Applied Statistics*, 32(2):127–140, 2005.

[205] D. Midthune, M. Fay, L. Clegg, and E. Feuer. Modeling reporting delays and reporting corrections in cancer registry data. *Journal of the American Statistical Association*, 100(469):61–70, 2005.

[206] R. Wasserstein and N. Lazar. The ASA's statement on p-values: context, process, and purpose, 2016.

[207] J. Lawless. Adjustments for reporting delays and the prediction of occurred but not reported events. *Canadian Journal of Statistics*, 22(1):15–31, 1994.

[208] P. Rosenberg. A simple correction of AIDS surveillance data for reporting delays. *J Acquir Immune Defic Syndr*, 3(1):49–54, 1990.

[209] P. Bacchetti, M. Segal, and N. Jewell. Backcalculation of HIV infection rates. *Statistical Science*, pages 82–101, 1993.

[210] R. Bellocco and I. Marschner. Joint analysis of HIV and AIDS surveillance data in back-calculation. *Statistics in Medicine*, 19(3):297–311, 2000.

[211] M. Barbosa and C. Struchiner. The estimated magnitude of AIDS in brazil: a delay correction applied to cases with lost dates. *Cadernos de Saúde Pública*, 18:279 – 285, 02 2002.

[212] S. Baker. The multinomial-poisson transformation. *The Statistician*, pages 495–504, 1994.

[213] P. Fader and B. Hardie. A note on modelling underreported poisson counts. *Journal of Applied Statistics*, 2000.

[214] R. Sakena and R. Kumar. Certain transformations of basic hypergeometric functions of two variables. *Le Matematiche*, 44(2):333–344, 1991.

[215] N. Temme. Large parameter cases of the Gauss hypergeometric function. *Journal of computational and applied mathematics*, 153(1):441–462, 2003.

[216] B. Sharma. A note on hypergeometric functions of two variables. *Indagationes Mathematicae (Proceedings)*, 79(2):169 – 172, 1976.

[217] J. Orbe, E. Ferreira, and V. Núñez-Antón. Comparing proportional hazards and accelerated failure time models for survival analysis. *Statistics in medicine*, 21(22):3493–3510, 2002.

[218] M. Rosen and J. Beck. *Human Immunodeficiency Virus and the Lung*. Lung Biology in Health and Disease. Taylor & Francis, 1998.

[219] D. Cox and G. Medley. A process of events with notification delay and the forecasting of AIDS. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 325(1226):135–145, 1989.

[220] G. Medley, V. Zunzunegui, R. Bueno, and D. Gai. The use of AIDS surveillance data for short-term prediction of AIDS cases in Madrid, Spain. *European journal of epidemiology*, 7(4):349–357, 1991.

Appendix A

HIV-AIDS European Case Definition

Since 1982, many definitions have been used for national and international reporting. In 2012, European Parliament and the Council established case definitions for AIDS and HIV on Commission Decision of 27/09/2012 to the Community network under Decision No 2119/98/EC .

- Clinical Criteria (AIDS)

  Any person who has any of the clinical conditions as defined in the European AIDS case definition for:

  - Adults and adolescents $\geq$15 years
  - Children < 15 years of age Laboratory Criteria (HIV)
  - Adults, adolescents and children aged $\geq$18 months At least one of the following three:
  - Positive result of a HIV screening antibody test or a combined screening test (HIV antibody and HIV p24 antigen) confirmed by a more specific antibody test (e.g. Western blot)
  - Positive result of 2 EIA antibody test confirmed by a positive result of a further EIA test
  - Positive results on two separate specimens from at least one of the following three:
  - Detection of HIV nucleic acid (HIV-RNA, HIV-DNA)
  - Demonstration of HIV by HIV p24 antigen test, including neutralisation assay
  - Isolation of HIV
  - Children aged < 18 months Positive results on two separate specimens (excluding cord blood) from at least one of the following three:
  - Isolation of HIV
  - Detection of HIV nucleic acid (HIV-RNA, HIV-DNA)
  - Demonstration of HIV by HIV p24 antigen test, including neutralisation assay in a child $\geq$ 1 month of age

Epidemiological Criteria: NA
Case Classification
A. Possible case: NA
B. Probable case: NA
C. Confirmed case:

- HIV infection: Any person meeting the laboratory criteria for HIV infection

- AIDS: Any person meeting the clinical criteria for AIDS and the laboratory criteria for HIV infection

ECDC case definition for adults and adolescents differs from the CDC's revised definition by not including the CD4 cell counts as an AIDS-defining criterion [218].

HIV disease staging and classification systems are critical tools for tracking and monitoring the epidemic. Two major classification systems currently are in use: the CDC Classification System and the WHO Clinical Staging and Disease Classification System [99, 100].

Both systems use the large spectrum of specific HIV - related clinical manifestations. While WHO system (revised in 2007) can be used readily in resource-constrained settings, CDC classification system (revised in 1993) assesses the severity of HIV by adding to the clinical manifestations the CD4 cell count. For CDC the definition of AIDS includes all HIV - infected individuals with CD4 counts of $< 200 cells/\mu L$ as well as those with certain HIV - related conditions and symptoms. Although the fine points of the classification system rarely are used in the routine clinical management of HIV infected patients, a working knowledge of the staging criteria is useful in patient care. In addition, the CDC system is used in clinical and epidemiology research. [99, 100].

Appendix B

European AIDS case definition

HIV/AIDS Surveillance in Europe
Surveillance du VIH/SIDA en Europe

Fourth Quarterly Report 1995

Quatrième rapport trimestriel 1995

EUROPEAN CENTRE FOR THE EPIDEMIOLOGICAL MONITORING OF AIDS
CENTRE EUROPEEN POUR LA SURVEILLANCE EPIDEMIOLOGIQUE DU SIDA
WHO-EC Collaborating Centre on AIDS - Centre Collaborateur OMS-CE sur le SIDA
94410 Saint-Maurice, France

European Case Definition for AIDS Surveillance in Children*

Revision 1995

The following revised European case definition for surveillance of acquired immunodeficieny syndrome (AIDS) in children is based on the 1987 revision of CDC/WHO case definition for AIDS [1, 2] and incorporates modifications consistent with the CDC 1994 revised classification system for human immunodeficiency virus (HIV) infection in children.[3] This European definition was prepared by the European Centre for the Epidemiological Monitoring of AIDS, Saint-Maurice, France following a decision taken by national representatives for AIDS surveillance in Europe and paediatric HIV/AIDS experts at a meeting held in Saint-Maurice on 12-13 January 1995. The main modifications concern changes in the diagnosis of HIV infection (*Annex 1*) and in the definitions of HIV encephalopathy and HIV wasting syndrome (*Annex 2*).

For surveillance purposes, a case of AIDS in a child* is defined as an illness characterized by one or more of the following "indicator" diseases, depending on the status of laboratory evidence of HIV infection, as shown below.

I. Without Laboratory Evidence for HIV Infection

If laboratory tests for HIV were not performed or gave inconclusive results (*Annex 1*) and the patient had no other cause of immunodeficiency listed in Section I.A below, then any disease listed in Section I.B indicates AIDS if it was diagnosed by a definitive method (*Annex 2*).

A. Causes of immunodeficiency that disqualify diseases as indicators of AIDS in the absence of laboratory evidence for HIV infection

1. high-dose or long-term systemic corticosteroid therapy or other immuno-suppressive/cytotoxic therapy ≤3 months before the onset of the indicator disease
2. any of the following diseases diagnosed ≤3 months after diagnosis of the indicator disease: Hodgkin's disease, non-Hodgkin's lymphoma (other than primary brain lymphoma), lymphocytic leukaemia, multiple myeloma, any other cancer of lymphoreticular or histiocytic tissue, or angioimmunoblastic lymphadenopathy
3. a genetic (congenital) immunodeficiency syndrome or an acquired immunodeficiency state atypical of HIV infection, such as one involving hypogammaglobulinaemia
4. exposure to abnormal levels of radiation

* Less than 13 years of age (Belarus, France and United Kingdom: less than 15 years)

Définition européenne pour la surveillance des cas de SIDA chez l'enfant*

Révision de 1995

La nouvelle définition européenne pour la surveillance des cas de syndrome d'immunodéficience acquise (SIDA) chez l'enfant est basée sur la définition CDC/OMS du SIDA modifiée en 1987 [1, 2] et tient compte de la révision de 1994 du système de classification du CDC de l'infection par le virus de l'immunodéficience humaine (VIH).[3] Cette nouvelle définition européenne a été préparée par le Centre Européen pour la Surveillance Epidémiologique du SIDA, Saint-Maurice, France, à la suite de la décision prise par les représentants nationaux de la surveillance du SIDA en Europe et par des spécialistes de l'infection à VIH et du SIDA chez l'enfant, lors d'une réunion qui a eu lieu à Saint-Maurice les 12 et 13 janvier 1995. Les principales modifications concernent le diagnostic de l'infection à VIH (*Annexe 1*) et les définitions de l'encéphalopathie due au VIH et du syndrome cachectique dû au VIH (*Annexe 2*).

Dans un but de surveillance, un cas de SIDA chez l'enfant* est défini par une ou plusieurs des pathologies indicatrices suivantes, en fonction de la connaissance du statut sérologique vis-à-vis du VIH.

I. En l'absence de preuve biologique de l'infection à VIH

Si les examens de laboratoire pour le VIH n'ont pas été faits ou si leur résultat est indéterminé (*Annexe 1*) et si l'enfant n'a pas une des causes d'immunodéficience énumérées ci-dessous dans le paragraphe I.A, chacune des pathologies énumérées dans le paragraphe I.B permet de porter le diagnostic de SIDA dans la mesure où le diagnostic de certitude de la pathologie a été établi (*Annexe 2*).

A. Causes d'immunodéficience qui empêchent de porter le diagnostic de SIDA lors du diagnostic d'une des pathologies énumérées en I.B

1. corticothérapie par voie générale, à dose élevée ou au long cours, ou autre traitement immunosuppresseur ou cytotoxique dans les trois mois précédant le diagnostic de la pathologie
2. chacune des maladies suivantes diagnostiquées dans les trois mois suivant le diagnostic de la pathologie : maladie de Hodgkin, lymphome non hodgkinien (autre que le lymphome cérébral primitif), leucémie lymphoïde, myélome multiple, autre cancer du système lymphoréticulaire ou du tissu histiocytaire, lymphadénopathie angio-immunoblastique
3. syndrome d'immunodéficience congénitale, ou immunodéficience acquise non évocatrice de l'infection à VIH telle que l'immunodéficience par hypogammaglobulinémie
4. exposition à des doses anormales de radiations

* Agés de moins de 13 ans (Belarus, France et Royaume-Uni : moins de 15 ans)

B. Indicator diseases diagnosed definitively (*Annex 2*)

1. candidiasis of the oesophagus, trachea, bronchi, or lungs
2. cryptococcosis, extrapulmonary
3. cryptosporidiosis with diarrhoea persisting >1 month
4. cytomegalovirus disease of an organ other than liver, spleen, or lymph nodes in a child >1 month of age
5. herpes simplex virus infection causing a mucocutaneous ulcer that persists longer than 1 month; or bronchitis, pneumonitis, or oesophagitis for any duration affecting a child >1 month of age
6. Kaposi's sarcoma
7. lymphoma of the brain (primary)
8. lymphoid interstitial pneumonia or pulmonary lymphoid hyperplasia (LIP/PLH complex)
9. *Mycobacterium avium* complex or *M. kansasii* disease, disseminated (at a site other than or in addition to lungs, skin, or cervical or hilar lymph nodes)
10. *Pneumocystis carinii* pneumonia
11. progressive multifocal leukoencephalopathy
12. toxoplasmosis of the brain affecting a child >1 month of age

II. With Laboratory Evidence for HIV Infection

Regardless of the presence of other causes of immunodeficiency (I.A), in the presence of laboratory evidence for HIV infection (*Annex 1 – paragraph A*), any disease listed above (I.B) or below (II.A or II.B) indicates a diagnosis of AIDS.

A. Indicator diseases diagnosed definitively (*Annex 2*)

1. serious bacterial infections, multiple or recurrent (i.e. any combination of at least two culture-confirmed infections within a 2-year period), of the following types: septicaemia, pneumonia, meningitis, bone or joint infection, or abscess of an internal organ or body cavity (excluding otitis media, superficial skin or mucosal abscesses, and indwelling catheter-related infections)
2. coccidioidomycosis, disseminated (at a site other than or in addition to lungs or cervical or hilar lymph nodes)
3. HIV encephalopathy
4. histoplasmosis, disseminated (at a site other than or in addition to lungs or cervical or hilar lymph nodes)
5. isosporiasis with diarrhoea persisting >1 month
6. lymphoma, small, noncleaved cell (Burkitt's), or immunoblastic or large cell lymphoma of B-cell or unknown immunologic phenotype
7. any mycobacterial disease caused by mycobacteria other than *Mycobacterium tuberculosis,* disseminated (at a site other than or in addition to lungs, skin, or cervical or hilar nodes)
8. *Mycobacterium tuberculosis,* disseminated or extrapulmonary
9. *Salmonella* (nontyphoid) septicaemia, recurrent
10. HIV wasting syndrome (emaciation, "slim disease")

B. Pathologies indicatrices diagnostiquées avec certitude (*Annexe 2*)

1. candidose de l'oesophage, de la trachée, des bronches, ou des poumons
2. cryptococcose extra-pulmonaire
3. cryptosporidiose avec diarrhée persistant depuis plus d'un mois
4. infection à cytomégalovirus atteignant au moins un organe autre que le foie, la rate ou les ganglions, chez un enfant âgé de plus de 1 mois
5. infection à virus herpès simplex entraînant un ulcère cutanéo-muqueux persistant plus d'un mois, ou infection herpétique bronchique, pulmonaire ou oesophagienne quelle que soit sa durée, chez un enfant âgé de plus de 1 mois
6. sarcome de Kaposi
7. lymphome cérébral (primitif)
8. pneumonie interstitielle lymphoïde ou hyperplasie pulmonaire lymphoïde (complexe PIL/HPL)
9. infection due au complexe *Mycobacterium avium* ou à *M. kansasii* disséminée (atteignant au moins un organe autre que les poumons, la peau ou les ganglions cervicaux ou hilaires)
10. pneumonie à *Pneumocystis carinii*
11. leucoencéphalopathie multifocale progressive
12. toxoplasmose cérébrale chez un enfant âgé de plus de 1 mois

II. En cas de preuve biologique de l'infection à VIH

Si l'infection à VIH a été diagnostiquée de façon formelle par le laboratoire (*Annexe 1 – paragraphe A*), toutes les pathologies énumérées ci-dessus (paragraphe I.B) ou ci-dessous (paragraphes II.A et II.B) permettent de porter le diagnostic de SIDA, *quelles que soient les autres causes d'immunodéficience (énumérées dans le paragraphe I.A).*

A. Pathologies indicatrices diagnostiquées avec certitude (*Annexe 2*)

1. infections bactériennes graves, multiples ou récidivantes (au moins deux infections confirmées par culture dans une période de 2 ans), sous une des formes suivantes: septicémie, infection pulmonaire, méningite, infection osseuse ou articulaire, abcès viscéral ou d'une cavité naturelle (à l'exception des otites moyennes, des abcès superficiels cutanés ou muqueux, et des infections sur cathéter)
2. coccidioïdomycose disséminée (atteignant au moins un organe autre que les poumons ou les ganglions cervicaux ou hilaires)
3. encéphalopathie due au VIH
4. histoplasmose disséminée (atteignant au moins un organe autre que les poumons ou les ganglions cervicaux ou hilaires)
5. isosporidiose avec diarrhée persistant depuis plus d'un mois
6. lymphome à petites cellules non clivées (type Burkitt), lymphome immunoblastique, ou lymphome à grandes cellules de type B, ou autre lymphome de phénotype immunologique inconnu
7. infection à mycobactéries autres que *Mycobacterium tuberculosis*, disséminée (atteignant au moins un organe autre que le poumon, la peau ou les ganglions cervicaux ou hilaires)
8. infection à *Mycobacterium tuberculosis*, disséminée ou extra-pulmonaire
9. septicémie récidivante à *Salmonella* autre que *S. typhi*
10. syndrome cachectique dû au VIH

B. Indicator diseases diagnosed presumptively (by a method other than those in *Annex 2*)

Note: Given the seriousness of diseases indicative of AIDS, it is generally important to diagnose them definitively (*Annex 2*), especially when therapy that would be used may have serious side effects or when definitive diagnosis is needed for eligibility for antiretroviral therapy. Nonetheless, in some situations, a patient's condition will not permit the performance of definitive tests. In other situations, accepted clinical practice may be to diagnose presumptively based on the presence of characteristic clinical and laboratory abnormalities. Guidelines for presumptive diagnoses are given in *Annex 3*.

1. candidiasis of the oesophagus
2. cytomegalovirus retinitis with loss of vision
3. Kaposi's sarcoma
4. lymphoid interstitial pneumonia or pulmonary lymphoid hyperplasia (LIP/PLH complex)
5. mycobacterial disease (acid-fast bacilli with species not identified by culture), disseminated (involving at least one site other than or in addition to lungs, skin, or cervical or hilar lymph nodes)
6. *Pneumocystis carinii* pneumonia
7. toxoplasmosis of the brain affecting a child >1 month of age

III. With Laboratory Evidence against HIV Infection

With laboratory test results negative for HIV infection (a non-reactive HIV antibody enzyme immunoassay [EIA] without a reactive or positive result on any other HIV detection test, if done) a diagnosis of AIDS for surveillance purposes is ruled out *unless*:

A. all the other causes of immunodeficiency listed above in Section I.A are excluded **and**

B. the child has had either:

1. *Pneumocystis carinii* pneumonia diagnosed by a definitive method (*Annex 2*)
   or
2. a) any of the other diseases indicative of AIDS listed above in Section I.B diagnosed by a definitive method (*Annex 2*)
   and
   b) evidence of immunodeficiency using appropriate age-specific CD4+ lymphocyte standards (see 1994 revised classification system for human immunodeficiency virus infection in children less than 13 years of age[3])

B. Pathologies indicatrices dont le diagnostic est présomptif (diagnostic par une méthode autre que celles énumérées dans l'*Annexe 2*)

Note : Etant donné la gravité des pathologies indicatrices de SIDA, il est important qu'elles soient diagnostiquées avec certitude (*Annexe 2*), en particulier lorsque les traitements préconisés peuvent entraîner des effets secondaires graves ou lorsque la confirmation du diagnostic est nécessaire pour l'instauration d'un traitement antirétroviral. Cependant, l'état du malade ne permet pas toujours d'effectuer les examens requis pour le diagnostic de certitude. Dans certains cas, un diagnostic de présomption peut être porté sur la présence de caractéristiques cliniques et d'anomalies biologiques. Les critères pour le diagnostic de présomption sont donnés dans l'*Annexe 3*.

1. candidose oesophagienne
2. rétinite à cytomégalovirus avec atteinte de la vision
3. sarcome de Kaposi
4. pneumonie interstitielle lymphoïde ou hyperplasie pulmonaire lymphoïde (complexe PIL/HPL)
5. infection à mycobactéries (bacilles acido-alcoolo-résistants non identifiés par la culture), disséminée (atteignant au moins un organe autre que les poumons, la peau ou les ganglions cervicaux ou hilaires)
6. pneumonie à *Pneumocystis carinii*
7. toxoplasmose cérébrale chez un enfant âgé de plus de 1 mois

III. En cas de preuve biologique de non infection par le VIH

En cas de résultats d'examens de laboratoire négatifs pour le VIH (négativité du test ELISA, ainsi que de tous les autres tests de détection du VIH s'ils ont été faits), le diagnostic de SIDA est porté, dans un but de surveillance, si :

A. toutes les autres causes d'immunodéficience énumérées dans le paragraphe I.A ont été exclues
et

B. l'enfant a :

1. une pneumonie à *Pneumocystis carinii* diagnostiquée avec certitude (*Annexe 2*)
   ou
2. a) une des autres pathologies énumérées dans le paragraphe I.B, diagnostiquée avec certitude (*Annexe 2*)
   et
   b) une immunodéficience diagnostiquée en utilisant les normes de lymphocytes CD4+ spécifiques par âge (voir la révision de 1994 du système de classification de l'infection VIH chez l'enfant de moins de 13 ans[3])

References

1. Centers for Disease Control and Prevention. 1987 revision of case definition for AIDS for surveillance purposes. Morbidity and Mortality Weekly Report, 1987;36 (1S):4-15.
2. World Health Organization. 1987 revision of CDC/WHO case definition for AIDS. Weekly Epidemiological Record, 1988; 63: 1-7
3. Centers for Disease Control and Prevention. 1994 revised classification system for human immunodeficiency virus infection in children less than 13 years of age. Morbidity and Mortality Weekly Report, 1994; 43 (RR-12): 1-10.

Annex 1. Diagnosis of human immunodeficiency virus (HIV) infection in children

A. HIV infected child

1. A child <18 months of age who is known to be HIV seropositive or born to an HIV- infected mother
   and either
   - has positive results on two separate determinations (excluding cord blood) from one or more of the following HIV detection tests:
     - HIV culture,
     - HIV polymerase chain reaction,
     - HIV antigen (p24),

     or
   - meets criteria for acquired immunodeficiency syndrome (AIDS) diagnosis according to the present AIDS surveillance case definition.

2. A child ≥18 months of age born to an HIV-infected mother or any child infected by blood, blood products, or other mode of transmission (e.g. sexual contact) who either
   - is HIV-antibody positive by repeatedly reactive enzyme immunoassay (EIA) and by a confirmatory antibody test
     or
   - meets any of the criteria in paragraph 1. above.

B. Perinatally-exposed — Infection status indeterminate

A child who does not meet the criteria above who either
- is HIV seropositive by EIA and by a confirmatory antibody test and is <18 months of age at the time of test
  or
- has unknown antibody status, but was born to a mother known to be infected with HIV.

C. Perinatally-exposed — Seroreverter

A child who is born to an HIV-infected mother and who
- has been documented as HIV-antibody negative (i.e. two or more negative EIA tests performed at 6-18 months of age or one negative EIA test after 18 months of age)
  and
- has had no other laboratory evidence of infection (has not had two positive viral detection tests, if performed)
  and
- has not had an AIDS indicator disease

Note: Passively acquired maternal HIV antibodies may be present in the child until 18 months of age. This may hamper the serological diagnosis of HIV infection in children vertically exposed to HIV infection.

Annexe 1. Diagnostic de l'infection par le virus de l'immunodéficience humaine (VIH) chez l'enfant

A. Enfant infecté par le VIH

1. Enfant âgé de moins de 18 mois, dont la sérologie vis-à-vis du VIH est positive ou qui est né d'une mère infectée par le VIH,
   et
   - qui a deux résultats positifs de détection du VIH lors de deux prélèvements différents (excluant les prélèvements du sang du cordon) pour au moins l'un des examens suivants :
     - culture virale VIH,
     - PCR VIH,
     - antigénémie VIH p24,

     ou
   - qui vérifie les critères de diagnostic du SIDA selon la présente définition du SIDA.

2. Enfant âgé d'au moins 18 mois, qui est né d'une mère infectée par le VIH ou qui a été infecté par transfusion ou par injection de dérivés sanguins ou par tout autre mode connu de transmission du VIH (par voie sexuelle par exemple), et
   - qui a une sérologie VIH positive par des tests ELISA répétées et par un test de confirmation
     ou
   - qui vérifie les critères mentionnés dans le paragraphe 1 ci-dessus.

B. Exposition périnatale – Infection par le VIH non déterminée

Enfant qui ne vérifie pas les critères cités plus haut et :
- qui a une sérologie VIH positive par un test ELISA et par un test de confirmation et qui est âgé de moins de 18 mois lors du test;
  ou
- qui a un statut sérologique inconnu, mais qui est né d'une mère infectée par le VIH.

C. Exposition périnatale – Enfant non infecté par le VIH

Enfant qui est né d'une mère infectée par le VIH et :
- qui a une sérologie VIH négative (au moins deux tests négatifs par ELISA entre l'âge de 6 et de 18 mois, ou un test négatif par ELISA après l'âge de 18 mois);
  et
- qui n'a pas d'autre preuve biologique d'infection par le VIH (n'a pas eu deux tests positifs de détection du virus s'ils ont été faits);
  et
- qui n'a pas eu de pathologies indicatrices du SIDA

Note : Les anticorps anti-VIH maternels (transmis à l'enfant à travers la barrière placentaire) peuvent persister chez l'enfant jusqu'à l'âge de 18 mois. La seule détection des anticorps anti-VIH chez un enfant âgé de moins de 18 mois qui a été exposé verticalement (mère-enfant) au VIH ne permet donc pas de porter le diagnostic d'infection à VIH chez cet enfant.

Annex 2. Definitive Diagnostic Methods for Diseases Indicative of AIDS in Children

Disease	Definitive Diagnostic Method
cryptosporidiosis isosporiasis Kaposi's sarcoma lymphoma lymphoid interstitial pneumonia *Pneumocystis carinii* pneumonia progressive multifocal leukoencephalopathy toxoplasmosis of the brain	microscopy (histology or cytology)
candidiasis	either by gross inspection by endoscopy or autopsy or by microscopy (histology or cytology) on a specimen obtained directly – not from a culture–from the tissues affected (including scrapings from the mucosal surface)
coccidioidomycosis cryptococcosis cytomegalovirus herpes simplex virus histoplasmosis	microscopy (histology or cytology), culture or detection of antigen in a specimen obtained directly from the tissues affected or a fluid from those tissues
tuberculosis other mycobacteriosis salmonellosis other bacterial infection	culture
HIV encephalopathy	**at least one** of the following progressive findings present for at least 2 months in the absence of a concurrent illness other than HIV infection that could explain the findings: - failure to attain or loss of developmental milestones or loss of intellectual ability, verified by standard developmental scale or neuropsychological tests - impaired brain growth or acquired microcephaly demonstrated by head circumference measurements or brain atrophy demonstrated by computerized tomography or magnetic resonance imaging (serial imaging is required for children <2 years of age) - acquired symmetric motor deficit manifested by two or more of the following: paresis, pathologic reflexes, ataxia, or gait disturbance

Annexe 2. Méthodes de diagnostic de certitudes des pathologies indicatrices de SIDA chez l'enfant

Pathologie	Méthode de diagnostic de certitude
cryptosporidiose isosporidiose leucoencéphalopathie multifocale progressive lymphome pneumonie interstitielle lymphoïde pneumonie à Pneumocystis carinii sarcome de Kaposi toxoplasmose cérébrale	microscopie (histologie ou cytologie)
candidose	examen macroscopique par endoscopie ou autopsie, ou examen microscopique (histologie ou cytologie) direct (et non sur une culture) d'un prélèvement de tissu infecté (y compris prélèvement des muqueuses)
coccidioïdomycose cryptococcose histoplasmose infection à cytomégalovirus infection à virus herpès simplex	microscopie (histologie ou cytologie), culture ou détection d'antigène sur un prélèvement obtenu directement du tissu infecté ou sur un liquide qui en est issu
tuberculose autre mycobactériose salmonellose autre infection bactérienne	culture
encéphalopathie due au VIH	**au moins** l'un des signes évolutifs suivants, constaté sur une période d'au moins deux mois, en l'absence de maladie autre que l'infection à VIH qui pourrait expliquer ces signes : - retard ou perte des acquisitions psychomotrices ou intellectuelles, évalués par des échelles de développement ou des tests neuropsychologiques adaptés - ralentissement de la croissance cérébrale ou microcéphalie acquise prouvés par la mesure du périmètre crânien, ou atrophie corticale prouvée par scanner ou IRM (plusieurs examens d'imagerie médicale sont nécessaires pour les enfants âgés de moins de 2 ans) - déficit moteur symétrique acquis, se manifestant par au moins deux des signes suivants : parésie, réflexes pathologiques, ataxie ou troubles de la marche

Annex 2. Definitive Diagnostic Methods for Diseases Indicative of AIDS in Children (continued)

HIV wasting syndrome	**at least one** of the following findings in the absence of a concurrent illness other than HIV infection that could explain the findings: - persistent weight loss >10% of baseline - downward crossing of at least two of the following percentile lines on a nationally accepted weight-for-age chart: 95th, 75th, 50th, 25th, 5th in a child ≥1 year of age - <5th percentile on a nationally accepted weight-for-height chart on two consecutive measurements, ≥30 days apart **and** either: - chronic diarrhoea (i.e. at least two loose stools per day for ≥30 days) **or** - documented fever (for ≥30 days, intermittent or constant).

Annexe 2. Méthodes de diagnostic de certitudes des pathologies indicatrices de SIDA chez l'enfant (suite)

syndrome cachectique dû au VIH	**au moins l'un des signes suivants**, en l'absence d'une maladie autre que l'infection à VIH qui pourrait expliquer ces signes : - perte de poids durable supérieure à 10 % du poids de base - perte de poids entraînant le franchissement par la courbe de poids selon l'âge d'au moins deux courbes correspondant aux 95ème, 75ème, 50ème, 25ème et 5ème percentiles calculés selon les normes établies pour le pays, chez un enfant âgé d'au moins 1 an - poids inférieur au 5ème percentile de la courbe de poids selon la taille établie selon les normes du pays, lors de deux pesées espacées d'au moins 30 jours; **et** - diarrhée chronique (au moins deux selles molles par jour pendant au moins 30 jours) **ou** - fièvre prouvée (continue ou intermittente, pendant au moins 30 jours)

Annex 3. Suggested Guidelines for Presumptive Diagnosis of Diseases Indicative of AIDS

Disease	Presumptive Criteria
candidiasis of oesophagus	- recent onset of retrosternal pain on swallowing **and** - oral candidiasis diagnosed by the gross appearance of white patches or plaques on an erythematous base or by the microscopic appearance of fungal mycelial filaments in an uncultured specimen scraped from the oral mucosa
cytomegalovirus retinitis	a characteristic appearance on serial ophthalmoscopic examinations (e.g. discrete patches of retinal whitening with distinct borders, spreading in a centrifugal manner, following blood vessels, progressing over several months, frequently associated with retinal vasculitis, haemorrhage and necrosis); resolution of active disease leaves retinal scarring and atrophy with retinal pigment epithelial mottling
mycobacteriosis	microscopy of a specimen from stool or normally sterile body fluids or tissue from a site other than lungs, skin, or cervical or hilar lymph nodes, showing acid-fast bacilli of a species not identified by culture
Kaposi's sarcoma	a characteristic gross appearance of an erythematous or violaceuous plaque-like lesion on skin or mucous membrane (presumptive diagnosis of Kaposi's sarcoma should not be made by clinicians who have seen few cases of it)
lymphoid interstitial pneumonia	bilateral reticulonodular interstitial pulmonary infiltrates present on chest X-ray for ≥2 months with no pathogen identified and no response to antibiotic treatment

Annexe 3. Diagnostic présomptif des pathologies indicatrices de SIDA

Pathologie	Critères diagnostiques
candidose oesophagienne	- douleur rétrosternale à la déglutition d'apparition récente, **et** - candidose buccale diagnostiquée macroscopiquement (plaques blanches sur une base érythémateuse) ou microscopiquement (filaments mycéliens à l'examen direct – et non sur une culture – d'un prélèvement de la muqueuse buccale)
rétinite à cytomégalovirus	aspect caractéristique sur les examens répétés du fond d'œil (taches rétiniennes blanchâtres, à bords nets, s'étendant de manière centrifuge en suivant les vaisseaux sanguins, évoluant sur plusieurs mois, accompagnées fréquemment de vascularite rétinienne avec hémorragies et nécrose) ; séquelles rétiniennes à type de cicatrices et d'atrophie avec taches épithéliales pigmentées
infection à mycobactéries	mise en évidence, à l'examen microscopique d'un prélèvement de selles, de liquides corporels normalement stériles ou de tissus provenant d'un organe autre que les poumons, la peau ou les ganglions cervicaux ou hilaires, de bacilles acido-alcoolo-résistants non identifiés par la culture
sarcome de Kaposi	plaques érythémateuses ou violacées caractéristiques à l'examen macroscopique de la peau ou des muqueuses (un diagnostic présomptif de sarcome de Kaposi ne doit pas être porté par des cliniciens qui n'ont vu que peu de cas)
pneumonie interstitielle lymphoïde	infiltrats pulmonaires interstitiels, réticulo-nodulaires et bilatéraux, à la radiographie pulmonaire, depuis plus de 2 mois, en l'absence de germes identifiés et en l'absence de réponse au traitement antibiotique

Annex 3. Suggested Guidelines for Presumptive Diagnosis of Diseases Indicative of AIDS (continued)

Pneumocystis carinii pneumonia	- a history of dyspnoea on exertion or nonproductive cough of recent onset (within the past 3 months) **and** - chest X-ray evidence of diffuse bilateral interstitial infiltrates or gallium scan evidence of diffuse bilateral pulmonary disease **and** - arterial blood gas analysis showing an arterial pO_2 of <70mm Hg or a low respiratory diffusing capacity (<80% of predicted values) or an increase in the alveolar-arterial oxygen tension gradient **and** - no evidence of a bacterial pneumonia
toxoplasmosis of the brain	- recent onset of a focal neurologic abnormality consistent with intracranial disease or a reduced level of consciousness **and** - evidence by brain imaging (computed tomography or nuclear magnetic resonance) of a lesion having a mass effect or the radiographic appearance of which is enhanced by injection of contrast medium **and** - serum antibody to toxoplasmosis or successful response to therapy for toxoplasmosis

Annexe 3. Diagnostic présomptif des pathologies indicatrices de SIDA (suite)

pneumonie à *Pneumocystis carinii*	- dyspnée d'effort ou toux non productive d'apparition récente (dans les trois mois précédents) **et** - infiltrats interstitiels diffus bilatéraux à la radiographie pulmonaire, ou aspect de pneumopathie diffuse bilatérale à la scintigraphie au gallium **et** - gaz du sang artériels avec une PaO_2 < 70 mm Hg, une faible capacité de diffusion (< 80% des valeurs prévues), ou une augmentation du gradient alvéolo-capillaire en O_2 **et** - absence de pneumopathie bactérienne
toxoplasmose cérébrale	- anomalies neurologiques focalisées d'apparition récente évoquant une lésion intra-cérébrale, ou troubles de la conscience **et** - mise en évidence, par scanner ou par IRM, d'une lésion entraînant un effet de masse ou prenant le produit de contraste **et** - sérologie de la toxoplasmose positive ou réponse au traitement de la toxoplasmose

Appendix C

Codes of AIDS Indicator Disease

1 = Bacterial infections, multiple or recurrent in a child under 13 years of age

2 = Candidiasis of bronchi, trachea, or lungs

3 = Candidiasis, oesophageal

4 = Coccidioidomycosis, disseminated or extrapulmonary

5 = Cryptococcosis, extrapulmonary

6 = Cryptosporidiosis, intestinal with diarrhoea (>1 months duration)

7 = Cytomegalovirus disease (other than liver, spleen, or nodes) in a patient over one month of age

8 = Cytomegalovirus retinitis (with loss of vision)

9 = Herpes simplex: chronic ulcer(s) (>1 months duration); or bronchitis, pneumonitis, or oesophagitis in a patient over one month of age

10 = Histoplasmosis, disseminated or extrapulmonary

11 = Isosporiasis, intestinal with diarrhoea (>1 months duration)

12 = Mycobacterium avium complex or M. kansasii, disseminated or extrapulmonary

13 = Mycobacterium tuberculosis, pulmonary in an adult or an adolescent (aged 13 years or over)*

14 = Mycobacterium tuberculosis, extrapulmonary

15 = Mycobacterium, other species or unidentified species, disseminated or extrapulmonary

16 = Pneumocystis carinii pneumonia

17 = Pneumonia, recurrent in an adult or an adolescent (aged 13 years or over)*

18 = Progressive multifocal leukoencephalopathy

19 = Salmonella (non typhoid) septicaemia, recurrent

20 = Toxoplasmosis of brain in a patient over one month of age

21 = Cervical cancer, invasive in an adult or an adolescent (aged 13 years or over)*

22 = Encephalopathy, HIV-related

23 = Kaposi s sarcoma

24 = Lymphoid interstitial pneumonia in a child under 13 years of age

25 = Lymphoma, Burkitt s (or equivalent term)

26 = Lymphoma, immunoblastic (or equivalent term)

27 = Lymphoma, primary, of brain

28 = Wasting syndrome due to HIV

30 = Opportunistic infection(s), not specified

31= Lymphoma(s), not specified

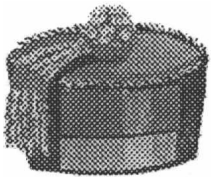
Appendix D

Portuguese Regulation

ANEXO II

Insígnias doutorais

Barrete



Capelo

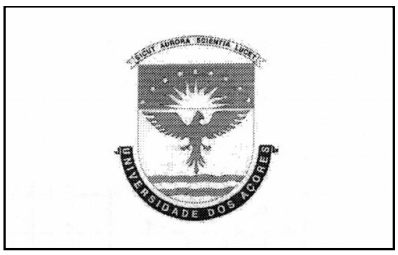
(Frente)

(Costas)

ANEXO III

Brasão de armas

Bandeira



ANEXO IV

Emblema

ANEXO V

Selo branco

ANEXO VI

Ex-líbris

MINISTÉRIO DA SAÚDE

Portaria n.º 258/2005

de 16 de Março

Em Portugal, a tabela de doenças de declaração obrigatória está ordenada de acordo com o código da 10.ª Revisão da Classificação Internacional de Doenças, conforme a deliberação n.º 131/97, de 27 de Julho, e constante da Portaria n.º 1071/98, de 31 de Dezembro.

Considerando que a monitorização e a projecção no curto e médio prazos da infecção por VIH é fundamental para a sua prevenção e controlo, o que apenas se torna exequível com o conhecimento do padrão epidemiológico da infecção do VIH em Portugal:

Manda o Governo, pelo Ministro da Saúde, o seguinte:

1.º A infecção pelo VIH passa a integrar a lista de doenças de declaração obrigatória, sendo por este meio alterada a tabela anexa à Portaria n.º 1071/98, de 31 de Dezembro.

2.º A declaração é obrigatória aquando do diagnóstico em qualquer estádio da infecção por VIH de portador assintomático (PA), complexo relacionado com a sida (CRS-LGP) e sida, e sempre que se verifique mudança de estadiamento ou óbito.

3.º É aprovado o modelo de folha de notificação relativa à vigilância epidemiológica da infecção por VIH, anexo a esta portaria e dela fazendo parte integrante.

4.º É revogada a Portaria n.º 103/2005, de 25 de Janeiro.

5.º A presente portaria produz efeitos desde a data da sua assinatura.

O Ministro da Saúde, *Luís Filipe da Conceição Pereira,* em 1 de Fevereiro de 2005.

Centro de Vigilância Epidemiológica das Doenças Transmissíveis

N.º / SIDA *
*A preencher pelo CVEDT

Vigilância Epidemiológica da Infecção pelo VIH

Folha de Notificação (ver instruções no verso, s.f.f.)

Ministério da Saúde

1. Tipo / classificação

SIDA ☐
CRS-LGP ☐
PA ☐

CDC*

	A	B	C
1			
2			
3			

* Se também disponível

2. Datas

Ano provável de infecção _____
Notificação ___/___/___ Diagnóstico ___/___/___
1.os Sintomas ___/___/___ Falecimento ___/___/___

3. Dados de codificação

Último apelido *(3 prim.as consoantes)* _____
Primeiro nome próprio *(2 prim.as consoantes)* _____
Sexo *(M/F)* ____ Data de nasc. ___/___/___ Idade _____
Naturalidade _____
Nacionalidade _____

4. Residência

Distrito _____ Concelho _____
País de resid.ª no provável contágio _____
País de resid.ª nos 1.os sintomas _____

5. Motivo

Motivo da consulta/internamento ou do teste _____
_____
_____

6. Gravidez

Gravidez à data de diagnóstico? SIM ☐ NÃO ☐
Categoria de transmissão da mãe nos casos de mãe-para-filho
☐ Toxicodependente IV
☐ Heterossexual
☐ Transfundida Data ___/___/___ País _____
☐ Outras/Indeterminada

7. Viagens/estadas no estrangeiro c/ possibilidades de contágio

País	Datas	Tipo de contágio
_____	___/___/___	_____
_____	___/___/___	_____
_____	___/___/___	_____
_____	___/___/___	_____
Serviço militar fora de Portugal		
_____	___/___/___	_____
_____	___/___/___	_____

8. Categorias de transmissão

☐ Bissexual ☐ Heterossexual
☐ Homossexual ☐ Toxicodep. IV
☐ Diálise renal ☐ Hemofílico tratado c/ concentrados
☐ Hemofílico tratado/crioprecipitados/plasma
☐ Infecção nosocomial
☐ Transfundido Data ___/___/___ País _____
☐ Transplantado Data ___/___/___
☐ Trab. sexo ☐ Transmissão mãe-para-filho
☐ Outras categorias (especificar) _____
_____

9. Características do parceiro no contacto heterossexual

☐ Desconhecido ☐ Hemofílico
☐ HIV 1 positivo ☐ HIV 2 positivo
☐ Homem Bissexual
☐ Originário/residente de país estrang. Qual? _____
☐ Trab. sexo ☐ Toxicodependente IV
☐ Transfundido ☐ Nenhum dos grupos mencionados

10. Doenças indicadoras de SIDA

1. Doença _____
   Método de diagnóstico _____
   Data ___/___/___ Serviço _____
2. Doença _____
   Método de diagnóstico _____
   Data ___/___/___ Serviço _____
3. Doença _____
   Método de diagnóstico _____
   Data ___/___/___ Serviço _____
4. Doença _____
   Método de diagnóstico _____
   Data ___/___/___ Serviço _____

11. Serologia VIH

	Data	Data 1.º teste VIH+
☐ Anti-VIH 1	___/___/___	___/___/___
Obs. _____		
☐ Anti-VIH 2	___/___/___	___/___/___
Obs. _____		
☐ Anti-VIH 1+VIH 2	___/___/___	___/___/___
Obs. _____		
☐ WBlot 1	___/___/___	___/___/___
Obs. _____		
☐ WBlot 2	___/___/___	___/___/___
Obs. _____		
☐ Antigénio	___/___/___	___/___/___
Obs. _____		
☐ Outros	___/___/___	___/___/___
Obs. _____		
_____		
_____		

12. Entidade que notifica

Nome _____
Serviço _____
Hospital _____

13. Outros serviços que contactam ou contactaram com o doente

_____
_____
_____
_____

Data ___/___/___

Assinatura _____

Vigilância Epidemiológica da Infecção pelo VIH
Folha de Notificação

Instruções para o preenchimento

■ Escrever legivelmente com letra de imprensa.

■ **Ponto 8** – ***Categorias de Transmissão*** –

- pode ser assinalada mais do que uma categoria de transmissão;
- a opção ***Outras categorias*** refere-se a qualquer modo de transmissão não mencionado anteriormente como, por exemplo, corte, picada involuntária por agulha ou contactos com líquidos orgânicos.

■ **Ponto 10** – deve seguir-se a ***"Definição de Casos de SIDA para Fins de Vigilância Epidemiológica, Revisão de 1993"*** (Doc. 77 do C.V.E.D.T./Comissão Nacional de Luta Conta a SIDA, Junho de 1994).

■ **Mais Informações em www.sida.pt**

Envio da Folha de Notificação

■ Enviar a Folha de Notificação para:

Instituto Nacional de Saúde
Centro de Vigilância Epidemiológica
das Doenças Transmissíveis
Av. Padre Cruz
1649 – 016 LISBOA

Tel. 217 519 200
Fax. 217 590 441

Despacho Normativo n.º 17/2005

O regime de codificação das embalagens dos medicamentos encontra-se consagrado no Despacho Normativo n.º 1/2003, de 15 de Janeiro, alterado pelo Despacho Normativo n.º 4/2004, de 16 de Janeiro, publicado na sequência da entrada em vigor do Decreto-Lei n.º 270/2002, de 2 de Dezembro, que cria o sistema de preços de referência para os medicamentos comparticipados pelo Estado, e pelo Despacho Normativo n.º 34/2004, de 25 de Junho, publicado no *Diário da República,* 1.ª série-B, n.º 170, de 21 de Julho de 2004, na sequência da entrada em vigor do Decreto-Lei n.º 81/2004, de 10 de Abril, que introduziu o artigo 5.º-A ao Decreto-Lei n.º 101/94, de 19 de Abril.

O Despacho Normativo n.º 4/2004 prevê no seu n.º 3 que serão aprovadas por despacho as especificações técnicas da codificação de medicamentos, mediante proposta do INFARMED.

A solução técnica a que se chegou passa pela simplificação dos dados impressos nas embalagens, o aumento da universalidade do sistema e a capacidade de aquisição automática de dados fixos e variáveis dos medicamentos.

O sistema de códigos de barras ora adoptado é o seguinte:

O actual Código de Barras 39, correspondente ao número de registo da apresentação do medicamento;

Um Código de Barras 39 complementar, alfanumérico, que, em conjunto com o anterior, permite o acesso a uma base de dados que contém, entre outros elementos, o lote, a validade e os preços.

Aproveita-se ainda a oportunidade para introduzir algumas modificações consideradas adequadas, nomeadamente a possibilidade de recolha de dados para suporte informático de elementos das especialidades farmacêuticas.

Appendix E

Portuguese Notification Form

Centro de Vigilância Epidemiológica das Doenças Transmissíveis

Vigilância Epidemiológica da Infecção pelo VIH

Folha de Notificação (ver instruções no verso, s.f.f.)

N.º / SIDA *

** A prencher pelo CVEDT*

Ministério da Saúde

1. tipo / classificação

SIDA ☐

CRS-LGP ☐

PA ☐

CDC*

	A	B	C
1			
2			
3			

* Se também disponível

2. datas

Ano provável de infecção _____

Notificação __aa__/__mm__/__dd__ Diagnóstico __aa__/__mm__/__dd__

1.os Sintomas __aa__/__mm__/__dd__ Falecimento __aa__/__mm__/__dd__

3. dados de codificação

Último apelido *(3 prim.as consoantes)* _____

Primeiro nome próprio *(2 prim.as consoantes)* _____

Sexo *(M/F)* _____ Data de nasc. __aa__/__mm__/__dd__ Idade _____

Naturalidade _____

Nacionalidade _____

4. residência

Distrito _____ Concelho _____

País de resid.a no provável contágio _____

País de resid.a nos 1.os sintomas _____

5. motivo

Motivo da consulta/internamento ou do teste _____

_____

_____

6. gravidez

Gravidez à data de diagnóstico? SIM ☐ NÃO ☐

Categoria de transmissão da mãe nos casos de mãe-para-filho

☐ Toxicodependente IV

☐ Heterossexual

☐ Transfundida Data __aa__/__mm__/__dd__ País _____

☐ Outras/Indeterminada

7. viagens/estadas no estrangeiro c/ possibilidades de contágio

País	Datas	Tipo de contágio
_____	__aa__/__mm__/__dd__	_____
_____	__aa__/__mm__/__dd__	_____
_____	__aa__/__mm__/__dd__	_____
_____	__aa__/__mm__/__dd__	_____

Serviço militar fora de Portugal

_____	__aa__/__mm__/__dd__	_____
_____	__aa__/__mm__/__dd__	_____

8. categorias de transmissão

☐ Bissexual ☐ Heterossexual

☐ Homossexual ☐ Toxicodep. IV

☐ Diálise renal ☐ Hemofílico tratado c/ concentrados

☐ Hemofílico tratado/crioprecipitados/plasma

☐ Infecção nosocomial

☐ Transfundido Data __aa__/__mm__/__dd__ País _____

☐ Transplantado Data __aa__/__mm__/__dd__

☐ Trab. sexo ☐ Transmissão mãe-para-filho

☐Outras categorias (especificar)_____

_____

9. características do parceiro no contacto heterossexual

☐ Desconhecido ☐ Hemofílico

☐ HIV 1 positivo ☐ HIV 2 positivo

☐ Homem Bissexual

☐ Originário/residente de país estrang. Qual? _____

☐ Trab. sexo ☐ Toxicodependente IV

☐ Transfundido ☐ Nenhum dos grupos mencionados

10. doenças indicadoras de SIDA

1. Doença _____
   Método de diagnóstico _____
   Data __aa__/__mm__/__dd__ Serviço _____
2. Doença _____
   Método de diagnóstico _____
   Data __aa__/__mm__/__dd__ Serviço _____
3. Doença _____
   Método de diagnóstico _____
   Data __aa__/__mm__/__dd__ Serviço _____
4. Doença _____
   Método de diagnóstico _____
   Data __aa__/__mm__/__dd__ Serviço _____

11. serologia VIH

	Data	Data 1.º teste VIH+
☐ Anti-VIH 1	__aa__/__mm__/__dd__	__aa__/__mm__/__dd__
Obs. _____		
☐ Anti-VIH 2	__aa__/__mm__/__dd__	__aa__/__mm__/__dd__
Obs. _____		
☐ Anti-VIH 1+VIH 2	__aa__/__mm__/__dd__	__aa__/__mm__/__dd__
Obs. _____		
☐ WBlot 1	__aa__/__mm__/__dd__	__aa__/__mm__/__dd__
Obs. _____		
☐ WBlot 2	__aa__/__mm__/__dd__	__aa__/__mm__/__dd__
Obs. _____		
☐ Antigénio	__aa__/__mm__/__dd__	__aa__/__mm__/__dd__
Obs. _____		
☐ Outros	__aa__/__mm__/__dd__	__aa__/__mm__/__dd__
Obs. _____		

12. entidade que notifica

Nome _____

Serviço _____

Hospital _____

13. outros serviços que contactam ou contactaram com o doente

_____

_____

_____

_____

Data ___/___/___

Assinatura _____

Appendix F

Literature Review

Over almost 40 years, researcher have been interested with obtaining valid estimates of recent incidence of reportable diseases which are, inevitable, subject to delays. Accurate and timely estimates are necessary for research evidence guiding Public Health decision-makers. A Literature Review summary table of key publications in reporting delay and the incidence of HIV-AIDS is given in F.1 .

Table F.1: Reporting Delays and HIV-AIDS Incidence - Modelling Approaches

Publication	Approach, Objectives and Estimation Technique	Application	Strengths [+] and Limitations [-]
Nonparametric Analysis of Truncated Survival Data, with Application to AIDS, 1988, Biometrika, [201]	Regression models based on reverse time hazards. Developed methods for estimating and comparing the identifiable aspects of the distribution for several groups. Non- parametric estimations based on conditional log-likelihood.	Reported to CDC infection and induction times from persons infected by contaminated blood transfusions and developed AIDS. Data was aggregated by grouping dates of infection ans AIDS into 3-month intervals.	[+] Nonparametric methods can estimate the induction distribution based on a identifiable portion. Can be generalized to allow a categorization of cases by the type of opportunistic illness. Can be used for comparing groups and stratified data. Can be used to estimate the cumulative distribution of time of infection. [-] No direct use of the nonparametric cumulative distribution for predict the future. Although it can be generalized to proportional hazards regression, it is difficult due to the complex dependence between the identifiable portion of the cumulative distribution on the model parameters and baseline hazard functions.
A Process of Events with Notification Delay and the Forecasting of AIDS, 1989, Philosophical Transactions of Royal Society [219]	Analysis and prediction of a point process in the presence of delay in the notification of the occurrence of an event. A parametric approach is taken to study the process and to predict future properties of the system. Application to AIDS data. Used several distributions for the reporting delay and presented one (weighted sum of two first order gamma distributions). Conditional log-likelihood	Case bases AIDS cases in the United Kingdom. Use an exponential model for the incidence function.	[+] Process formulated in continuous time [-] Epidemic is not described adequately by simple exponential growth. Since the number of cases are inflated by the long tail reporting delay distribution, delay lag were arbitrarily truncated in 3 years long.
Reporting delays and the Incidence of AIDS, 1990, Journal of the American Statistical Association [121]	Assume Y_{tu} as independent and with Poisson distribution. Describe the joint probability function and assume the likelihood function can be written as a product of the conditional likelihood and the marginal. Describe methods for estimating both the probability distribution of reporting delays and the actual incidence of AIDS. Are described categorical, mixed categorical / continuous -time models for reporting delay. Categorical and continuous-time model for incidence of AIDS. Estimates Incidence of HIV based on back calculation. Maximum Likelihood	Reported AIDS cases to the CDC. Aggregated cases of AIDS stratified by region. Estimate incidence of HIV among non-drug-using homosexual men based on backcalculation.	[+] Describes the models for reporting delay and incidence under the assumption of Poisson distribution. The corresponding likelihood and the decomposition between conditional and marginal likelihood. Can accommodate both truncated and censored data. Estimates of HIV incidence consistent with serological data. [-] Imposes constraints on the distributions that may be satisfied only by some type of regression models. The counts of cross-classified HIV - AIDS cases are very unlikely to be independent Poisson. If the cases are reported in batches the independent multinomial assumption does not capture the pattern. Restrictiveness of assumption that incidence and reporting delay models share no common parameters. Do not reflect uncertainty in the HIV incubation density which is assumed to be stationary.
Statistical modelling of the AIDS Epidemic for forecasting Health Care Needs, 1990, Biometrics [202]	Extend the method of back-calculation to allow for a multistage model. Fitting regression models based on reverse time hazard. Quasi-likelihood estimates through iteratively reweighted least squares under weakly parametric models.	AIDS Epidemic in the United States with CD4 cells count.	[+] Short term projections are given of both AIDS incidence and the numbers of HIV-infected incorporating CD4 cells count. Consider individuals stratified by stage of disease. Evaluate the impact of therapeutics. [-] Uncertainty in the hazard functions of disease progression, in the parametric model for the infection rate, in the AIDS incidence data, in the efficacy of treatment, and in the proportions of the HIV-infected individuals receiving treatment. Backcalculation methods assume a progressive disease model. Data most be transformed.

(Table continues...)

Table F.1: Reporting Delays and HIV-AIDS Incidence - Modelling Approaches

Publication	Approach, Objectives and Estimation Technique	Application	Strengths [+] and Limitations [-]
A simple correction of AIDS Surveillance Data for Reporting Delays, 1990, Journal of Acquired Immune Deficiency Syndromes [208]	A simple noniterative method based on Poisson model counts. Maximum likelihood estimations.	CDC cross-classified surveillance data.	[+] Easily implemented [-] Assume reporting delay is stationary.
Regression Models for Right Truncated Data With Applications to AIDS Incubation Times and Reporting Lags, 1991, Statistica Sinica [122]	Fitting regression models based on a reverse time hazard function to obtain nonparametric estimates of AIDS latency distribution. The reverse time hazards are expressed for both discrete and continuous times. Discuss regression models and related semi-parametric methods for right truncated data. Develop tests concerning the independence of time and truncation time. Full log-likelihood in the discrete case or the partial in the continuous case.	AIDS cases reported to the CDC, with diagnoses data and reporting date. AIDS cases on risk group "Blood transfusion" since the date of infection is known and age of patients.	[+] The reverse time hazards are expressed for both discrete and continuous times. The models allow the easy comparison of truncated distributions for different populations. May be used to assess the quasi-stationarity of lag distributions which arise in connection with pair of events. The models provide flexible and realistic representations of covariate effects for truncated data. [-] Can be applied only to the proportional hazards model and accommodate only right truncated data. Assume that the reporting delay is stationary.
The use of AIDS surveillance data for short term predictions of AIDS cases in Madrid, Spain, 1991, European Journal of Epidemiology [220]	Join estimation of incidence function and distribution of reporting delays. Preliminary attempt to predict the numbers of AIDS cases in the Community of Madrid up to 1992. Uses surveillance data to determine trends and predict the likely future patterns of diagnoses. A parametric approach is taken to study the process and to predict future properties of the system. Conditional likelihood	Surveillance data of Community of Madrid, with dates of diagnosis, report, alive and deaths due to AIDS.	[+] Gives approximate confidence limits for the number of new diagnoses. The parametric form here should capture the broad shape, and is acceptable to produce estimates. A starting point for planning. [-] Just an application. No new methods are described. The confidence limits are skewed since underreporting was not included. Prediction depend completely on the choice of the incidence function.
Regression Analysis of Censored and truncated Data: Estimating Reporting-Delay Distributions and AIDS Incidence from Surveillance Data, 1994, International Biometrics Society [8]	Truncation is a time of missing data. Discuss fitting regression models to data that can be truncated and even censored in arbitrary intervals. Expectation-Maximization as a data augmentation and account truncation.	CDC surveillance data grouped based on quarterly intervals and stratified by regions. Apply two high risk-groups: men who have sex with men and male intravenous drug users.	[+] Allows to model the time trends of the reporting delay separately for each of the regions. Proved string interaction between the chronologic time trends and geographic regions. Permits fitting a variety of models, including those for categorical data analysis , to arbitrarily censored and truncated data using standard software packages. [-] Analysis conditional on the cases with delays less than 6 years. Small number of parameters model.
Backcalculation of HIV Infection Rates, 1993, Statistical Science [209]	Reviews of the backcalculation technique, focusing on the key assumptions of the method, including the necessary information regarding incubation, reporting delay, and models for the infection curve. A summary is given of the extent to which the appropriate external information is available and whether checks of the relevant assumptions are possible through use of data on AIDS incidence from surveillance systems. New features of the approach include incorporation of seasonal variation in diagnosis rates, smooth nonparametric estimation of both the HIV infection curve and nonstationary aspects of the incubation period and reporting delay distributions, and an analysis of residuals from backcalculation fits. A likelihood approach to backcalculation.	AIDS incidence data in the United States with diagnosis and reporting dates, transmission group, incubation time.	[+] New features of the approach include incorporation of seasonal variation in diagnosis rates, smooth nonparametriec estimation of both the HIV infection curve and nonstationary aspects of the incubation period and reporting delay distributions, and an analysis of residuals from backcalculation fits. Unexplained lack of fit is examined and discussed. [-] Uncertainty associated with backcalculation estimates caused by misspecified assumptions and inaccurate external estimates of key components of the technique. Such uncertainty limits the usefulness of backcalculation and highlights the need for complementary approaches. Use of cohorts data which may not be representative of the more general population.
Adjustments for reporting delays and the prediction of occurred but not reported events, 1994, Canadian Journal of Statistics [207]	Estimate the number of events that have occurred but not yet been reported. Allow random temporal fluctuations in reporting delays, and consequently, confidence or prediction limits.	AIDS case from Canada	[+] The objective of this paper was to present methods of adjustment for reporting delays when delay probabilities vary somewhat over time, but not in a systematic fashion. [-] Need the investigator to choose the degree of smoothing. Discrete time framework.
A neural network for survival data, 1995, Statistics in Medicine [197]	Modelling censored survival data using the input-output relationship associated with a Feed - Forward neural network as the basis for a non-linear proportional hazard model for censored survival data. Maximum likelihood.	Survival data of men with prostatic carcinoma who entered a four-group clinical trial comparing the effect of several concentrations of diethylstilbestrol. Data on twelve covariates, beside the treatment group, were available, in addition to survival time and survival status.	[+] Introduce a broad applicability of the neural networks approach to survival data and any generalized linear models. Emphasis given to predictive power rather than inference about model parameters since the latter are generally difficult to interpret. It allows the topology of a single hidden layer feed-forward network to be used for representing the relationship between the hazard function and covariates. Back.propagation can be used for training censored survival data. [-] Comparisons of neural network models to other models in a variety of applications is necessary to evaluate the utility of this approach. The confidence interval must be constructed via bootstrap, which is computationally intensive. There are not available methods for computation of asymptotic variance for the statistics due to the functional dependences of the predicted outcomes on the survivals throughout the estimated parameters. A boostrap technique is necessary to obtain a more precise estimate of predictive accuracy for each model. Careful attention to computational issues is necessary due to the possible non unimodal likelihood function.

(Table continues...)

Table F.1: Reporting Delays and HIV-AIDS Incidence - Modelling Approaches

Publication	Approach, Objectives and Estimation Technique	Application	Strengths [+] and Limitations [-]
Using surveillance data to monitor trends in the AIDS epidemic, 1998, Statistics in Medicine [120]	Describes adjustments that accounts for delays in reporting AIDS cases, the lack of HIV-exposure information for some cases, and future diagnoses of AIDS-defining opportunistic illnesses among persons reported with AIDS under severe immunosuppression criteria of the 1993 AIDS surveillance case definition. Partition of the cases being analysed into clusters that are small enough for within-cluster reporting delays to be fairly homogeneous but large enough to yield precise estimates of their reporting delay distributions. First select the variables throughout binary response regression models and then the reporting delay is estimated separately for each group of cases defined by the cross-classification of levels identified during the first stage, and groups with similar estimated delay distributions are combined Conditional maximum likelihood.	CDC AIDS cases with diagnosis and reporting date, AIDS defining illness,State of residency at the time of AIDS diagnosis; risk group; race / ethnicity; indicators of whether the population state of residence at diagnosis exceeds one million and whether AIDS diagnosis and death occur during the same month	[+] Deterministic evaluation of the model by a CDC expect. Reporting delay stratified by variables. Determine a key determinant of the length of reporting delay that is the location and type of health care facility. Determine that the time variation of reporting delay is due to time variation of reporting behaviour of the reporting health-facility ('catch-up' periods). Takes the context of the reporting facility into consideration. Allow variations in AIDS incidence trends by mode of exposure to HIV, sex, race/ethnicity and geographic region which are in qualitative agreement with reported variations in HIV prevalence rates and changes in HIV prevalence over time. [-] Analysis of a short period of time. Reporting delay does not vary over six-year period within groups for which separate estimates are made. A maximum of four years of delay is assumed. Uncertainty in the exposure distributions. Underreporting was not covered.
Applications of multiple imputation in medical studies: from AIDS to United States National Health and Nutrition Examination Survey (NHANES) , 1999, Statistical Methods in Medical Research [116]	This paper reviews three applications of Rubin's method that are directly relevant for medical studies: estimating the reporting delay in acquired immune deficiency syndrome (AIDS) surveillance systems for the purpose of estimating survival time after AIDS diagnosis; missing data and noncompliance in randomized experiments, where a school choice experiment is used as an illustration; and handling nonresponse in United States National Health and Nutrition Examination Surveys . The emphasis of our review is on the building of imputation models (i.e. the first step), which is the most fundamental aspect of the method.	CDC surveillance data with diagnosis and reporting dates	[+] Its flexibility in separately handling the incomplete-data problems and the substantive analysis. In the context of dealing with nonresponse in public-use data files, it is a method without serious competition in terms of both generality and validity because of the unavoidable separation of the creators and the users of the database and because of the information and resource constraints on the average users for sensibly handling the nonresponse. When used for other purposes, it is an effective addition to a statistician's toolkit because of the conceptual and implementation simplicity offered by the separation of the tasks of handling incomplete data and analysing complete data. [-] If the imputation model is seriously flawed in terms of capturing the missing-data mechanism, then so will be any analysis based on such imputations. Less efficient than joint modelling of the missing-data mechanism and the substantive analysis, in terms of both statistical efficiency (e.g. avoiding the reliance on a finite number of imputations) and computational efficiency (e.g. avoiding simulation via the use of the EM algorithm).
Joint Analysis of HIV and AIDS Surveillance Data in Back - Calculation, 2000, Statistics in Medicine [210]	Generalization of back-calculation method of HIV cases based on AIDS surveillance data by incorporating information gained by linkage with an HIV surveillance system, containing data on the first positive HIV test. It allows a flexible HIV testing rate model, which incorporates dependence on both calendar time and since infection. Use a flexible approach to smoothing, based on generalized additive models. Em algorithm with generalized additive model smoothing.	Data of AIDS and HIV surveillance systems from Veneto, Itlay	[+] Generalization of backcalcualtion for inclusion of first HIV test data on AIDS-free individuals as well as those with AIDS. Estimates based on combination of data are more precise that based only on AIDS. The improvement in efficiency of HIV incidence estimates is consistent with results of other publications. Allows the assessment of trends in the testing rates of HIV positive individuals. [-] Estimates dependent on assumptions of incubation distribution. Needs to be extended for inclusion of CD4 cells count.
Data and projections of HIV and AIDS in Portugal, 2000, Journal of Applied Statistics [203]	Backcalculation method of back-calculation for obtaining forecasts of the portuguese surveillance system accounting for reporting delay and . The method used to estimate the reporting delay distribution is based on Poisson regression and involves cross-classifying each reported case by calendar time of diagnosis and reporting delay. The adjusted AIDS incidence data are then used to obtain short-term projections and lower bounds on the size of the AIDS epidemic. The estimation procedure 'back-calculates' from AIDS incidence data using the incubation period distribution to obtain estimates of the numbers previously infected. The expectation-maximization algorithm is used to obtain maximum-likelihood estimates when the density of infection times is parametrized as a step function.	Portuguese Surveillance data with diagnosis and reporting delay dates and risk groups.	[+] The methodology is applied to AIDS incidence data in Portugal for four different transmission categories: injecting drug users, sexual transmission (homosexual/bisexual and heterosexual contact) and other, mainly haemophilia and blood transfusion related. the method does not require knowledge of the total number of seropositives in the population. The projections indicate that the AIDS epidemic constitutes a major public health problem in Portugal. [-] It produces a lower bound, since it only estimates the cumulative number of individuals that will eventually develop AIDS from those already infected with HIV1. The method requires accurate data on the incidence of the disease over time, as well as information on the incubation period distribution. Uncertainty in the the incubation period distribution. Assumptions that activities (such as underreporting) are the same for all transmission categories or constant over time. Parametric hazard models for incubation or clinical latency period distribution with Weibull distribution.
Data and Projections of HIV / AIDS Cases in Portugal: An Unstoppable Epidemic?, 2005, Journal of Applied Statistics [204]	Back-calculation for estimating the size of Portuguese HIV epidemic. The EM algorithm was applied to obtain maximum likelihood estimates of the HIV incidence. The density of infection times was parametrized as a step function.	AIDS incidence in Portugal for four different transmission categories (injecting drug users, heterosexual, homo/bisexual and other) to obtain short-term projections (2002–2005) and an estimate of the minimum size of the epidemic. Reporting delay distribution estimated based on a conditional likelihood .	[+] Projections consistent with the observed trend of the epidemic. [-] Recent discussions suggest that the backcalculation is gradually becoming less appropriate for reliable incidence and prevalence estimates, as it does not take into account the effect of treatment. Lower bound of the epidemic defined. Incubation period distribution is not precisely known, there is inaccuracy in the observed disease incidence over time, the assumptions made relative to the underreporting percentage and treatment effect is not been taken into consideration.

(Table continues...)

Table F.1: Reporting Delays and HIV-AIDS Incidence - Modelling Approaches

Publication	Approach, Objectives and Estimation Technique	Application	Strengths [+] and Limitations [-]
Modelling Reporting Delays and Reporting Corrections in Cancer Registry Data, 2005, Journal of the American Statistical Association [205]	The primary objective is to predict the eventual net count, ideally after an infinite delay, based on the data collected up until the current reporting year. A secondary interest is in the reporting delays and corrections themselves, for data quality control purposes. We would like to know, for example, the average reporting delay, the percentage of diagnosed cancers reported within 2 years, the percentage of reporting errors in the data, and whether these measures have improved or worsened in recent years. Divide the SEER population into subpopulations based on registry and the usual subgroups used for reporting, that is, one for each combination of levels of the following variables: registries, year of diagnosis (17 years), gender, race, and 5-year age groups then assumed Poisso n counts. Conditional Maximum Likelihood.	SEER cancer data Set from US	[+] Explicitly jointly model the reporting corrections, model the delay distribution with very general models, combining aspects of previous nonparametric-like models with more parametric models and allow random reporting-year effects in the model. [-] All of the reporting models assume that the reporting process is relatively stable . The modified AIC model selection criterion performed well in simulations, but the authors noted that some care must be used when fitting nonstationary models, because there is a risk of confounding diagnosis year, reporting year, and delay time effects, which could have a significant effect on prediction of eventual counts and might not be detected by the AIC. It was e introduced a random reporting-year effects version of the reporting model that fit our data much better than did the nonrandom model, but still exhibited some lack of fit. For simplicity, it was assumed that the random effects were normally distributed and independent; alternatively, one could allow the random effects to be correlated or use nonparametric methods to estimate the distribution of effects. This model allows correlation within reporting year and it would be useful to also allow correlation with diagnosis year.
Modelling reporting delays for outbreak detection in infectious disease data, 2015, Journal of the Royal Statistical Society [198]	The main aim is to describe the reporting delay hazard, with the aim of gaining a better understanding of the reporting process, and to investigate whether temporal factors, notably calendar time, season and recent incidence, influence reporting delays. Use a continuous time spline-based model for the hazard of the delay distribution, along with an associated proportional hazards model. The delay distributions are found to have extremely long tails, the hazard at longer delays being roughly constant, suggestive of a memoryless process.	Public Health England LabBase surveillance database with 12 Infectious diseases: *Acinetobacter baumanii*, *Campylobacter jejuni*, *Chlamydia sp*, *Giardia lamblia*, *Norovirus* , *Salmonella abony*, *Salmonella braenderup* , *Salmonella brandenburg*, *Salmonella enteritidis* , *Salmonella infantis*, *Salmonella senftenberg*, *Salonella typhimurium*, Factors: calendar time, season and recent incidence .	[+] The modelling framework that we have used is based on a semiparametric regression model for the hazard. This allow to visualize the hazard in continuous time, and to study the effect of covariates in a natural fashion, with a simple relative hazard interpretation. Describe and characterize the delay distribution in its entirety, including long delays, which are often dropped from the analysis of surveillance data. [-] The major disadvantage of the approach is that it is more cumbersome than the generalized linear model method. Estimates of the hazard at long delays are sensitive to irregularities in the data, notably those resulting from discontinuities in reporting.
Two artificial neural networks for modelling discrete survival time of censored data, 2015, Journal Advances in Artificial Intelligence [186]	Present a comprehensive comparison of two different approaches of utilizing Artificial Neural Networks (ANN) in modelling smooth conditional hazard probability function. Use ANN in survival time modelling of skin cancer.	Male and Female melanoma cancer from Surveillance, Epidemiology, and end results program (SEER) data .	[+] Applying ANN is more applicable and efficient, especially when the data does not satisfy Cox PH assumptions. ANN does not require any assumptions that need to be justified, and it is more precise in fitting nonlinear models. [-] More investigation and statistical data analysis are required to better understand the causes of the differences between young males and females and to plan new strategies to fight the major pernicious form of skin cancer.
Models for surveillance data underreporting delay: applications to US veteran first-time suicide attempters, 2015, Journal of Applied Statistics [163]	Allow separate models for the incidence and reporting delay in a distribution-free fashion, but with joint inference for both modelling components, based on functional response models. Discuss inference about projections of future disease incidence to help identify significant shifts in temporal trends modelled based on the observed data.	Simulated and real data from US veteran first-time suicide attempters surveillance system.	[+] Allows modelling disease incidence and reporting delay separately to allow for accommodation of a subsample of data for modelling the reporting delay; Use of the Functional Response Models to frame the two modelling components (disease incidence and reporting delay) within the context of a single model; Use of a set of weighted generalized Estimating Equations adapted to the Functional Response Models to provide consistent parameter estimates and valid inference and presented new methods for projecting disease incidence. Simulated data provided good performance, even for relatively small sample sizes. [-] For modelling incidence assumption on distributions are needed

Appendix G

French Notification Form

Version 2015

ARS de :

Fiche clinicien : infection à VIH chez l'adulte (15 ans et plus)

Maladie à déclaration obligatoire (Art. L3113-1, R3113-1, R3113-2, R3113-3, R3113-5, D3113-7 du Code de la santé publique).

Critères de notification : toute sérologie VIH confirmée positive chez un sujet de 15 ans et plus.

Information individuelle des personnes : droit d'accès et de rectification (loi du 6 janvier 1978 modifiée) pendant 12 mois, par l'intermédiaire du médecin déclarant, auprès de l'Institut de veille sanitaire.

cerfa
12221*xx
50959#xx

Caractéristiques sociodémographiques

Code d'anonymat* : ☐☐☐☐ ☐☐☐☐ ☐☐☐☐ ☐☐☐☐ **Année de naissance :** ☐☐☐☐

Sexe* : ☐ H ☐ F ☐ transgenre H vers F ☐ transgenre F vers H

Pays de domicile* : ☐ France, préciser le département : ☐☐ ☐ autre pays

Lieu de naissance : ☐ France métropolitaine ☐ Guadeloupe, Martinique, Guyane, Saint-Martin, Saint-Barthélemy ☐ France d'outre-mer autre
☐ autre pays, préciser : . . . . . **Année d'arrivée en France :** ☐☐☐☐

Nationalité : ☐ française ☐ étrangère, préciser : . . . . . .

Profession (exercée ou non)* : . . . . . **En activité professionnelle :** ☐ oui ☐ non (formation, chômage, retraite, etc.)

Statut vital : ☐ vivant ☐ décédé, préciser la date du décès : ☐☐ ☐☐ ☐☐☐☐ ☐ inconnu

Sérologie VIH actuelle

Date du prélèvement de la sérologie positive : ☐☐ ☐☐ ☐☐☐☐

Motif(s) de réalisation de la sérologie actuelle* *(plusieurs choix possibles)* **:**

- ☐ risque d'exposition au VIH datant de moins de 6 mois
- ☐ risque d'exposition au VIH datant de 6 mois ou plus
- ☐ symptômes/signes cliniques liés au VIH
- ☐ infection sexuellement transmissible
- ☐ contrôle d'une séropositivité connue
- ☐ confirmation d'un TROD positif
- ☐ confirmation d'un autotest positif
- ☐ grossesse
- ☐ dépistage sans notion d'exposition au VIH
- ☐ autre, préciser : . . . . . .

Initiative de la sérologie actuelle* : ☐ proposée par le médecin ☐ demandée par la personne

Sérologies VIH antérieures et recours à une PrEP*

Négative : ☐ oui ☐ non ☐ inconnu
(y compris tests rapides)* Si oui, date de la dernière négative : ☐☐ ☐☐ ☐☐☐☐

Confirmée positive* : ☐ oui ☐ non ☐ inconnu
(si oui, merci de compléter néanmoins cette déclaration) Si oui, date de la première confirmée positive : ☐☐ ☐☐ ☐☐☐☐

Recours à une PrEP : ☐ oui ☐ non ☐ inconnu
(quelle que soit la période)

VHC, VHB et IST (syphilis, gonococcie, chlamydiose, LGV)

Ac anti-VHC au moment du diagnostic VIH : ☐ positif ☐ négatif ☐ inconnu

Ag HBs au moment du diagnostic VIH : ☐ positif ☐ négatif ☐ inconnu

IST au moment du diagnostic VIH ou dans les 12 mois précédents :
☐ oui, la(es)quelle(s) : . . . . . ☐ non ☐ inconnu

Statut clinique et immunologique au moment du diagnostic d'infection à VIH

Stade clinique : ☐ primo-infection symptomatique* ☐ asymptomatique ☐ symptomatique non sida ☐ inconnu
☐ sida*, préciser : - pathologie(s) inaugurale(s) de sida : . . . . . .
- date de diagnostic du sida : ☐☐ ☐☐ ☐☐☐☐
- traitements antirétroviraux pré-sida ≥3 mois : ☐ oui ☐ non ☐ inconnu
(au cours des deux dernières années)

Nombre de CD4 : ☐☐☐☐ /mm³ Date : ☐☐ ☐☐ ☐☐☐☐
(mesure la plus proche possible du diagnostic d'infection à VIH)

Charge virale : ☐☐☐☐☐ copies/ml Date : ☐☐ ☐☐ ☐☐☐☐
(mesure la plus proche possible du diagnostic d'infection à VIH)

Circonstances probables de la contamination

Mode(s) probable(s) de contamination :

- ☐ rapports homosexuels*, préciser s'il s'agit d'un homme ayant aussi des rapports hétérosexuels : ☐ oui ☐ non ☐ inconnu
- ☐ rapports hétérosexuels
- ☐ usage de drogues injectables
- ☐ autre, préciser : . . . . . .
- ☐ inconnu

La date de contamination peut-elle être estimée : ☐ oui ☐ non
Si oui, préciser à quelle date ou période : . . . . . .
et comment *(prise de risque unique, test antérieur...)* : . . . . . .

Lieu probable de contamination : ☐ France ☐ autre pays ☐ inconnu

Partenaire sexuel à l'origine probable de la contamination*

Préciser s'il s'agit d'un partenaire :

- ☐ homosexuel/bisexuel
- ☐ usager de drogues injectables
- ☐ originaire d'Afrique subsaharienne
- ☐ originaire des Caraïbes/Guyane
- ☐ appartenant à une autre population exposée au VIH
  Préciser : . . . . . .
- ☐ n'appartenant pas à une autre population exposée au VIH
  Préciser : . . . . . .
- ☐ aucune information disponible sur le partenaire

Surveillance virologique*

Après avoir été informée, la personne refuse les tests de la surveillance virologique : ☐ *(ne cocher cette case qu'en cas de refus de la personne)*

Observations

. . . . . .
. . . . . .

Biologiste à l'origine du diagnostic VIH

Nom : . . . . . Tél. : ☐☐ ☐☐ ☐☐ ☐☐ ☐☐

Courriel : . . . . . .

Nom et adresse du laboratoire (ou cachet) :

. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .

Date de notification : ☐☐ ☐☐ ☐☐☐☐

Clinicien déclarant

Nom : . . . . . Tél. : ☐☐ ☐☐ ☐☐ ☐☐ ☐☐

Courriel : . . . . . .

Cadre d'exercice :

- ☐ Hôpital ou clinique (nom) : . . . . . .
  Chef de service et spécialité : . . . . . .
- ☐ Médecine libérale, préciser la spécialité : . . . . . .
- ☐ Autre, préciser : . . . . . .

Adresse : . . . . . .
. . . . . .

Code postal : ☐☐☐☐☐ Ville : . . . . . .

Date de notification : ☐☐ ☐☐ ☐☐☐☐

Cette fiche de déclaration obligatoire doit être adressée au médecin de l'ARS sous pli portant la mention « secret médical ».

** Voir au verso de la fiche*

Ce formulaire téléchargeable n'est à utiliser que si vous ne parvenez pas à déclarer en ligne sur : www.e-do.fr

1 - Critères de notification de l'infection à VIH chez l'adulte (15 ans et plus)

Doit être déclaré tout diagnostic d'infection à VIH (**sérologie VIH confirmée positive**) chez une personne de 15 ans et plus.

2 - Mode de remplissage du formulaire par le clinicien

Caractéristiques sociodémographiques

- Le **code d'anonymat** est calculé directement dans l'application de déclaration en ligne (www.e-do.fr). À défaut, il peut être calculé au moyen du logiciel d'anonymisation fourni par l'InVS sur demande, à partir du prénom, de l'initiale du nom, de la date de naissance et du sexe de la personne. Ce code, indispensable à l'enregistrement du cas, permet de repérer les doubles déclarations.

- Concernant le sexe, le terme « **transgenre** » désigne toute personne dont l'identité de genre, l'expression ou le comportement sont en inadéquation avec son sexe biologique.

- Le **pays de domicile** correspond au lieu où vit habituellement la personne et non à une adresse temporaire liée à une prise en charge médicale ou autre.

- La **profession** ou la catégorie socioprofessionnelle (ex : cadre, profession intermédiaire, employé…) doit être précisée, que la personne soit en activité ou non. Si la personne n'est plus en activité, indiquer la dernière profession exercée. « Sans profession » concerne seulement les personnes n'ayant jamais exercé de profession.

Sérologie VIH actuelle : Les variables « **Motif(s) de réalisation de la sérologie actuelle** » et « **Initiative de la sérologie actuelle** » ont pour objectifs de décrire les circonstances de réalisation d'une sérologie VIH, dans le cadre des évolutions de la stratégie de dépistage.

Sérologies VIH antérieures et recours à une prophylaxie pré-exposition au VIH (PrEP)

- L'information sur l'existence de **sérologies antérieures négatives** et la date de la dernière négative est très importante, car indispensable à l'estimation de l'incidence du VIH (nombre annuel de nouvelles contaminations). Les tests rapides négatifs (TROD ou autotests) sont également à mentionner dans cette rubrique.

- La notion d'une **sérologie antérieure confirmée positive** est demandée, car la séropositivité peut être déjà connue sans jamais avoir été déclarée. Ne pas tenir compte des TROD ou autotests positifs, non confirmés par une sérologie classique.

Statut clinique et immunologique au moment du diagnostic d'infection à VIH

- La **primo-infection symptomatique** correspond à une période de réplication virale intense, au cours de laquelle la personne peut présenter des manifestations cliniques débutant 2 à 6 semaines après sa contamination (syndrome pseudo-grippal, pharyngite, éruption cutanée, adénopathies superficielles…).

- Si la personne est au **stade sida** au moment de la découverte de séropositivité, il est possible d'indiquer sur ce formulaire la(les) pathologie(s) inaugurale(s) de sida, la date de diagnostic du sida et la notion de traitements antirétroviraux. Dans ce cas, il est inutile de remplir un formulaire de déclaration de sida.

Circonstances probables de la contamination : Concernant les hommes contaminés par rapports homosexuels, l'intérêt de préciser s'ils ont aussi des rapports hétérosexuels est de caractériser la population bisexuelle découvrant sa séropositivité VIH et d'appréhender le risque de diffusion de l'infection à la population hétérosexuelle.

Partenaire sexuel à l'origine probable de la contamination

- Le libellé « **partenaire originaire des Caraïbes/Guyane** » inclut notamment la Guadeloupe et la Martinique.

- Les « **autres populations exposées au VIH** » incluent notamment les partenaires originaires d'une autre zone géographique à forte prévalence (Asie, Europe de l'Est…) et les partenaires prostitué(e)s.

Surveillance virologique : Il s'agit de tests complémentaires effectués par le Centre national de référence (CNR) du VIH sur le premier prélèvement réalisé pour le diagnostic, notamment du test d'infection récente qui permet de calculer l'incidence du VIH. Il vous incombe d'informer la personne séropositive sur cette surveillance virologique, et sur le fait qu'elle peut s'y opposer. Cochez l'item correspondant seulement en cas de refus.

Coordonnées du clinicien déclarant et du biologiste : Il vous est demandé d'indiquer vos coordonnées, ainsi que celles du biologiste à l'origine du diagnostic VIH, afin de permettre à l'InVS, pendant une période de 12 mois, de valider éventuellement certaines informations (code d'anonymat invalide, information manquante, feuillet médical non reçu…).

3 - Envoi du formulaire

Ce formulaire de déclaration obligatoire doit être adressé au médecin de l'ARS de votre lieu d'exercice, sous pli portant la mention « secret médical ». L'ARS transmettra ensuite les déclarations de sa région à l'InVS pour saisie et analyse des données.

Il vous est demandé de conserver un double de cette déclaration pendant 12 mois, pour permettre une validation éventuelle par le médecin de l'InVS.

Les notices d'information, destinées aux personnes dont vous déclarez l'infection VIH, sont disponibles sur le site web de l'InVS : www.invs.sante.fr

Appendix H

Italian Notification Form

SCRIVERE IN STAMPATELLO

84102

Cognome | Nome | Sesso M F

Comune di nascita (o stato estero) | Prov. | Data di nascita gg mm aa

Nazionalità | Comune di residenza | Prov.

da staccare e spedire in buste separate

SCRIVERE IN STAMPATELLO

Non compilare riservato ISS

Codice ISS

84102

Grado di istruzione: Nessuno | Licenza elementare | Licenza media | Diploma | Laurea

Diagnosi di AIDS effettuata in sede di autopsia: SI | NO

Se deceduto specificare:

Data decesso gg mm aa | Comune decesso | Prov.

Decesso associato all'AIDS: SI NO

Cause: iniziale | intermedia | terminale

INFORMAZIONI GENERALI

MALATTIE INDICATIVE DI IMMUNODEFICIENZA ACQUISITA

Diagnosi accertata

	Mese	Anno
1 - Candidosi - bronchi, trachea o polmoni		
2 - Candidosi esofagea		
3 - Carcinoma cervicale invasivo (non valido per caso pediatrico)		
4 - Coccidioidomicosi disseminata		
5 - Criptococcosi extrapolmonare		
6 - Criptosporidiosi intestinale cronica		
7 - Cytomegalovirus, malattia sistemica		
8 - Cytomegalovirus, retinite		
9 - Encefalopatia da HIV		
10 - Herpes simplex: ulcera/e croniche (>1 mese) o bronchite, polmonite o esofagite		
11 - Isosporidiosi cronica intestinale (>1 mese)		
12 - Istoplasmosi disseminata		
13 - Leucoencefalopatia Multifocale Progressiva		
14 - Linfoma di Burkitt (o equivalente)		
15 - Linfoma immunoblastico (o equivalente)		
16 - Linfoma primitivo cerebrale		
Micobatteriosi disseminata o extrapolmonare da:		
17 - M. tubercolosis		
18 - M. avium e M. kansasii		
19 - M. da altre specie o da specie non identificata		
20 - Polmonite da Pneumocystis carinii		
21 - Polmonite ricorrente (non valido per caso pediatrico)		
22 - Sarcoma di Kaposi		
23 - Sepsi da salmonella, ricorrente		
24 - Toxoplasmosi cerebrale		
25 - Tubercolosi polmonare (non valido per caso pediatrico)		
26 - Wasting Syndrome da HIV		

Diagnosi presuntiva

	Mese	Anno
2 - Candidosi esofagea		
8 - Cytomegalovirus, retinite		
Micobatteriosi disseminata o extrapolmonare da:		
17 - M. tubercolosis		
18 - M. avium e M. kansasii		
19 - M. da altre specie o da specie non identificata		
20 - Polmonite da Pneumocystis carinii		
21 - Polmonite ricorrente (non valido per caso pediatrico)		
22 - Sarcoma di Kaposi		
24 - Toxoplasmosi cerebrale		
25 - Tubercolosi polmonare (non valido per caso pediatrico)		

Malattie valide solo per i casi pediatrici (≤ 12 anni)

27 - Infezioni batteriche ricorrenti		
28 - Polmonite interstiziale linfoide		28 - Polmonite interstiziale linfoide

SCHEDA DI NOTIFICA DI SINDROME DA IMMUNODEFICIENZA ACQUISITA

Centro Operativo AIDS (COA) Viale Regina Elena, 299 - 00161 Roma - Tel. 06 49902584/2337 Fax 06 49902755 e-mail coa@iss.it

SCRIVERE IN STAMPATELLO

MODALITÀ DI TRASMISSIONE

Le modalità di trasmissione di seguito indicate derivano: dall'anamnesi del paziente ☐ dalla valutazione del medico ☐

1 - Rapporti omosessuali/bisessuali ☐
2 - Uso droghe e.v. ☐
3 - Somministrazione fattori coagulazione ☐
4 - Trasfusione ☐
Data trasfusione ☐☐ ☐☐ (mm aa)
Accertata positività trasfusione NO ☐ SI ☐
5 - Rapporti eterosessuali ☐
6 - Pre-perinatale ☐
7 - Rischio non determinato ☐
8 - Altro (vedi nota fattori rari) ☐
(specificare)

- Informazioni relative al rischio del partner o della madre
a - Rapporti omosessuali/bisessuali ☐
b - Uso droghe e.v. ☐
c - Somministrazione fattori coagulazione ☐
d - Trasfusione ☐
e - Prostituta/o ☐
f - Non noto ☐
g - Rapporti eterosessuali ☐
h - Altro (specificare)

- Positività HIV nota SI ☐ NO ☐

- Paese di provenienza: Italia ☐ Altro

Positività degli anticorpi anti-HIV1 SI ☐ NO ☐ Non nota ☐
Positività degli anticorpi anti-HIV2 SI ☐ NO ☐ Non nota ☐

DATI CLINICI

Dati clinici pre-AIDS

	Mese	Anno	
Data ultimo test HIV-negativo	☐☐	☐☐	
Data primo test HIV-positivo	☐☐	☐☐	
Data prima misurazione CD4	☐☐	☐☐	/mmc
e relativo valore assoluto CD4	☐☐☐☐		/mmc

Ha mai iniziato la terapia antiretrovirale prima della diagnosi di AIDS:

☐ Si ☐ No ☐ Non noto

Se **SI** indicare la terapia più aggressiva effettuata:

☐ Mono ☐ Duplice ☐ Tripla ☐ Non nota ☐ Altro

Profilassi infezioni opportunistiche SI ☐ NO ☐ Non noto ☐

Dati clinici alla diagnosi di AIDS

Valore assoluto dei CD4 ☐☐☐☐ /mmc

Viremia plasmatica ☐ Si ☐ No ☐ In corso

Metodo (specificare)

Cut-off ☐☐☐☐ (copie/ml)

Risultato ☐☐☐☐☐☐☐☐ (copie/ml)

SEGNALAZIONE

SCRIVERE IN STAMPATELLO

Medico che segnala il caso

Divisione/Clinica/Reparto

Ente/Ospedale di appartenenza

Indirizzo Ente/Ospedale via Città

CAP Prov. ☐☐ Tel. ☐☐☐☐ ☐☐☐☐☐☐☐☐

e-mail Fax ☐☐☐☐ ☐☐☐☐☐☐☐☐

Data compilazione scheda ☐☐ ☐☐ ☐☐ (gg mm aa)

NOTE

Non compilare riservato all'ISS

Data arrivo scheda ☐☐ ☐☐ ☐☐ (gg mm aa)

Rev. 7/2002 ISS

(copia per il COA, ISTITUTO SUPERIORE DI SANITÀ - LEB 35 - Viale Regina Elena, 299 - 00161 Roma)

Appendix I

Polish Notification Form

Pieczątka podmiotu wykonującego działalność leczniczą:	**ZLK-4** **Zgłoszenie podejrzenia lub rozpoznania zakażenia HIV/ zachorowania na AIDS/ zgonu osoby zakażonej HIV/ chorej na AIDS***	Adresat: **Państwowy Powiatowy/Graniczny* Inspektor Sanitarny** w .....

Resortowy kod identyfikacyjny podmiotu[1)]

Część I. Numer księgi rejestrowej

Część II. TERYT siedziby

Część III. Podmiot tworzący[2)]

Część IV. Specjalność komórki organizacyjnej

Uwagi:

[1)] Wypełnić zgodnie z rozporządzeniem Ministra Zdrowia z dnia 17 maja 2012 r. w sprawie systemu resortowych kodów identyfikacyjnych oraz szczegółowego sposobu ich nadawania (Dz. U. poz. 594).

[2)] Wypełnić w przypadku podmiotu leczniczego niebędącego przedsiębiorcą.

*Niepotrzebne skreślić.

I . ROZPOZNANIE/PODEJRZENIE*

1. Kod ICD-10 – **2. Określenie słowne** **3. Data** (dd/mm/rrrr) / /

4. Informacje dotyczące rozpoznania

A. Nowo wykryte zakażenie ludzkim wirusem niedoboru odporności (HIV)

Badanie potwierdzające:

numer badania ..... data badania : .....

metoda:

☐ western-blot ☐ badanie wirusologiczne ☐ badanie molekularne

☐ inna (wpisać jaka) .....

Jednostka zlecająca badanie: .... tel. .....

Laboratorium potwierdzające: .... tel. .....

Stan kliniczny przy rozpoznaniu HIV:

☐ ARS ☐ zakażenie bezobjawowe ☐ objawy choroby HIV, nie-AIDS ☐ AIDS

B. Rozpoznanie AIDS

Choroby wskazujące na AIDS rozpoznane metodą definitywną:

1. .....
2. .....
3. .....
4. .....

Choroby wskazujące na AIDS rozpoznane metodą prawdopodobną:

1. .....
2. .....
3. .....
4. .....

5. Zgon osoby zakażonej HIV/chorej na AIDS* - przyczyna podejrzenia lub rozpoznania zgonu

Kod ICD-10 – **Określenie słowne** **Data** (dd/mm/rrrr) / /

* Niepotrzebne skreślić.

II. DANE CHOREGO*

1. Nazwisko

2. Imię **3. Data urodzenia (dd/mm/rrrr)** **4. Nr PESEL****

5. Inny krajowy nr identyfikacyjny** **6. Płeć** (M, K) **7. Wiek*****

Adres miejsca zamieszkania:

8. Kod pocztowy **9. Miejscowość**

10. Ulica **11. Nr domu** **12. Nr lokalu**

* Jeżeli pacjent zastrzeże dane osobowe, należy wpisać: zamiast nazwiska i imienia - w poz. 1 i 2 inicjały imienia i nazwiska lub hasło, zamiast daty urodzenia - w poz. 7 wiek, zamiast adresu - w poz. 9 nazwę powiatu właściwego ze względu na miejsce zamieszkania. Pacjent nie może zastrzec danych o płci.

** Wypełnić w przypadku, gdy osobie nie nadano numeru PESEL, wpisując serię i numer paszportu albo numer identyfikacyjny innego dokumentu, na podstawie którego jest możliwe ustalenie danych osobowych.

*** Nie wypełniać w przypadku wpisania daty urodzenia.

III. DANE EPIDEMIOLOGICZNE

Badanie w kierunku HIV:

data ostatniego ujemnego badania: …….. … data pierwszego dodatniego badania: ……

AIDS: data rozpoznania pierwszego epizodu: ……

Liczba komórek CD4 w odstępie 3 miesięcy od rozpoznania HIV/AIDS: ……...

Leczenie ARV przed rozpoznaniem AIDS/zgonem:

☐ TAK ☐ NIE ☐ Nie wiadomo

Przynależność do grup szczególnie narażonych:

☐ IDU ☐ MSM,WSW ☐ CSW

☐ Pracownicy ochrony zdrowia/personel medyczny

Drogi nabycia zakażenia:

☐ kontakt homoseksualny ☐ kontakt heteroseksualny ☐ kontakt seksualny, brak dokładnych danych

☐ wstrzyknięcie narkotyku ☐ wertykalna ☐ transfuzja krwi/preparatów krwiopochodnych

☐ jatrogenna ☐ inna (jaka?) ……...

Inne choroby przenoszone drogą płciową: ……..

IV. DANE ZGŁASZAJĄCEGO LEKARZA/FELCZERA/KIEROWNIKA SZPITALA

1. Pieczątka imienna ..... 2. Telefon kontaktowy: ..... 3. Podpis .....

Appendix J

The Netherlands Patient Data Collection

PATIENT DATA COLLECTION

Patient Entry

Upon patient entry into the Stichting HIV Monitoring database, the following information is collected:

Items collected upon initial enrolment for HIV-infected adults		
Demographic data	Date of birth, gender, first and second nationality, country of birth, height, location of testing and health care body that referred pt to specialist	
History of HIV infection	Date of the last negative HIV-1 and HIV-2 test	
	Date of the first positive HIV-1 and HIV-2 test	
	Was the patient diagnosed with a primary HIV infection? (yes, no, most likely)	
	Primary HIV infection based on: recent negative test, indeterminant western-/immunoblot, patient knows moment of risk behaviour, symptoms reported < 6 months, other, unknown.	Symptoms reported < 6 months: Lymphadenopathy, flu like symptoms, fatigue, malaise, headache, fever, night sweats, photophobia, mouth ulcers, pharyngitis, neck stiffness, anorexia, nausea, vomiting, weight loss (>2,5 kg), diarrhoea, genital ulcers, anal ulcers, arthralgia / joint pain, myalgia / muscle pain, rash, confusion, gingivitis, unkown.
HIV transmission	The most likely transmission route: homosexual heterosexual injecting drug use (IDU) blood and blood products during pregnancy/partum via breastfeeding other and unknown	For sexual transmission, the most likely transmission route is entered: either a steady sexual partner or multiple sexual contacts
	Country where the patient became infected	
Congenital	Has the patient haemophilia?	
Intoxication	Data on smoking, alcohol consumption and drug intake	

Additional data for HIV-infected children	
Demographic data	Nationality and country of birth of patient's parents
Family data	HIV status of patient's mother, father, brothers and sisters
Perinatal data	Pregnancy duration, way of birth, weight at birth, Apgar scores, congenital defects, perinatal exposure to antiretroviral (ARV) therapy and co-medication, antenatal complications

Additional data for HIV-infected pregnant women		
Demographic data	Nationality and country of birth of patient's parents Patient's ethnicity (*'Asian', 'Caucasian', 'Black', 'other', or 'unknown)*	
Screening	Was the patient found to be HIV-positive at the national pregnancy screening?	
Visits to Gynaecologist	Visit date, Blood pressure	
Obstetric data	Nationality and country of birth biological father of child Expected birth date Number of pregnancies *(>37 weeks, between; 33-36 weeks, 28-32 weeks, 16-27 weeks)* Number of spontaneous abortions *(<16 weeks)*	Number of Apla Number of EUG Pregnancy risk Smoking during pregnancy Alcohol use during pregnancy Drugs use during pregnancy
Delivery data	Has there been a delivery/abortion? Date of delivery/abortion Sex of the baby Duration of pregnancy Child number Prophylactic antibiotics? Intra-uterine infection Duration of dilation Duration of ruptured membranes of expulsion	Mode of delivery Caesarean section? Fetal scalp electrode Episiotomy or rupture Birth weight of the baby Apgar scores after 1 minute/5 minutes Duration of stay in the incubator Perinatal mortality Breast-feeding?
Complications during pregnancy	Complications during and/or after birth? Blood loss during the first half of pregnancy? Blood loss during second half of pregnancy? Intercurrent infection? Version (attempt) with breech presentation? Pre-eclampsia?	Intra-uterine retardation of growth (sonography<p5%)? PPROM (preterm premature rupture of outer membranes) at how many weeks? Abdominal trauma at how many weeks?
Extra lab results	HIV1_DNA HIV2_DNA CMV_RNA EBV_RNA Anti-VCA-IGG Anti-VCA-IGM Anti-EA_IGG	Anti_EA_IGM Anti_EBNA_IGG T_Gondii_IGG T_Gondii_IGM Anti_VZV_IGG Anti_VZV_IGM Total protein urine (g/L)

After patient enrolment, clinical data are collected every time the patient is seen by his or her treating physician. These data contain the following information:

<table>
<tr><th colspan="3">Items collected at every follow-up visit for HIV-infected adults</th></tr>
<tr><td>Clinical examination</td><td colspan="2">Weight, blood pressure</td></tr>
<tr><td>CDC events
Start and stop date and the status of event at current visit (ongoing: yes or no).</td><td colspan="2">HIV-related events as classified by CDC. Definition of diagnosis (possible, presumptive or definitive) are recorded by standard protocol</td></tr>
<tr><td rowspan="2">Adverse events
Start and stop date and the status of event at current visit (ongoing: yes or no).</td><td colspan="2">Every event that results in a change of antiretroviral treatment is collected. In addition, the following events are always recorded:</td></tr>
<tr><td>Peripheral neuropathy
Myopathy
Lactate acidosis
Hepatic fibrosis / cirrhosis
Osteopenia / Osteoporosis
Hepatic steatosis
Hepatic encephalopathy
Oesophagus varices
Bleeding from the oesophagus varices
Hepatorenal syndrome
Liver transplantation
Pancreatitis
Nephrolithiasis
Renal insufficiency and failure
Kidney dialysis
Kidney transplantation
Lipodistrophy, fat loss in extremities
Lipodistrophy, central fat accumulation
Rash
Abacavir hypersensitivity
Sexual dysfunction (loss of libido, erectile dysfunction)
Congestive heart failure (cardiac decompensation</td><td>Non-AIDS malignancies
Anal dysplasie
Diabetes mellitus
Myocardial infarction
Hypertension
Arrhythmia
Cardiomyopathy
Stroke
Coronary artery by-pass grafting
Coronary angioplasty / stenting
Carotid endarterectomy
Pregnancy
Hospital admission
Liver cirrhosis
Liver fibrosis
Neurosyfilis
Pneumonia
Myopathy
Castle Mann disease
Salmonella sepsis
Herpes Zoster
Angina pectoris
Fractures pathological, specify
Fractures traumatic, specify</td></tr>
</table>

Antiretroviral therapy, (including hepatitis B treatment hepatitis C treatment) *Start units, route of admission, reason for stop and the status of medication at current visit (ongoing: yes or no t and stop date, dosage and)*	Standard stop reasons are as follows: Virological failure Immunological failure Patient's decision Toxicity New CDC-B and or CDC-C events Interaction with co-medication Simplification of the regimen Related to blood concentration of ARV Structured treatment interruption Newly available medication As a precaution Viral load undetectable Palliative / terminal Died	 Pregnancy wish Pregnancy End of pregnancy Compliance problems Other, specify Unknown Doses increase Start hepatitis treatment Hepatitis; completed treatment Hepatitis; no RVR (rapid viral response) Hepatitis; no EVR (early viral response) Hepatitis; no response (not specify) Protocol Viral load undetectable/ viral load decrease
Co-medication *Start and stop date and the medication status at current visit (ongoing: yes or no)*	CDC events, prophylaxis CDC events, treatment Anti-epileptic agents Anti-coagulant agents Platelet aggregation inhibitors Anti-hypertensive agents Anti-arrhythmic agents Lipid lowering agents Anti-diabetic agents Insulin and its derivatives	Anabolic steroids and appetite stimulants Medication that interacts with antiretroviral therapy Miscellaneous: megestrol acetate, dranabinol and methadone Chemotherapy
Lab results	*HIV virology: RNA* Value (copies/ml), laboratory, sample date, VL assay type, sample material, cut-off and undetectable: yes or no	
	Immunology: T-cell count Value, units, laboratory and sample date for the following determinates: CD4 count, CD8 count, CD4 percentage, CD8 percentage, CD4/CD8 ratio	

	Chemistry Value, units, laboratory and sample date for the following determinates: Glucose >N* Triglycerides always collected Amylase >250 mmol/l Cholesterol always collected ALAT/SGPT>3 x N* Cholesterol HDL always collected ASAT/SGOT>3 x N* Albumin always collected Alkaline phosphatase >3 x N* Bilirubin always collected Gamma GT >3 x N* Phosphate always collected Lactate>N* LDH always collected Creatinine always collected PTT always collected * N is normal value; can vary for different laboratories
	Haematology Value, units, laboratory and sample date for the following determinates: Haemoglobin <5.5 mmol/l Leukocytes <2.0 10e9/l Thrombocytes always collected
	Other viral infections Value (positive or negative), laboratory, sample date for the following determinates: HBsAg, HBsAb, HBcAb, HBeAg, HBeAb, HBV-DNA (quantitative and qualitative values), HCV-Ab, HCV-RNA (quantitative and qualitative values), CMV-IgG, CMV-IgM
	Sexually transmitted diseases Value, units, laboratory and sample date for the following determinates: Chlamydia Condylomata accuminata Gonorrhoea Human Papilloma virus Syphilis Herpes HSV 1,2 HTLV
	ART drug concentrations Plasma concentration, laboratory, sample data, time after drug intake, dosage and units of the medication
Patient's participation in clinical trials	Trial name, start and stop date, informed consent.
Additional data; Liver diagnostic (HIV patient and HIV patient with and hepatitis)	
Fibro scan	Results; date, which hospital fibro scan was done, how many measurement were done, how many successful measurements were done, Inter Quartile Range IQR, Median score 'stiffness', metavir range / metavir score (F0-F4), success range (%)
Liver pathology	Results; date, hepatocellular carcinoma, cirrhoses or fibroses, hepatic steatosis, HAI index, Ishak score, knodell score, metavir range / metavir score (F0-F4)
Radiology	Results: date, what radiology diagnostic was done: echo, CT or MRI, height/length/size spleen. focal lesions, portal flow/hypertension, cirrhoses or fibroses, hepatic steatosis, hepatocellular carcinoma, splenomegaly, ascitis, collateral vessels

Additional data for HIV and Hepatitis (HBV or HCV) –infected adults	
Vaccinations	HAV 1, HAV 2, HBV 1, HBV 2, HBV 3
HCC treatment	Start date, stop date, type of treatment
Viral response	Results: date, early viral response, rapid viral response, Sustained viral response.
Mutations and genotype	Results: date, (host cell) HBV genotype / mutations, (host cell) HCV genotype/mutations.
Extra Lab results	HBV-DNA, HCV-Ab, HCV-AG/AS, HDAg, HAV_Total, HAV_IGG, HAV_IGM

Additional data for HIV-infected children		
Clinical examination	Skull circumference, puberty stage	
CDC events *Start and stop date and the status of event at current visit (ongoing: yes or no).*	HIV-related events as classified by CDC. Definitions of diagnosis (possible, presumptive or definitive) are recorded by standard protocol. In addition to CDC-B and –C events, CDC-A events are also collected.	
Adverse events	Pathologic and traumatic fractures, abnormalities of psychological development, abnormalities of locomotion development, abnormalities of puberty development	
Additional treatment *Start and stop date, status at current visit*	Psychologist, pedagogue, psychiatrist, speech therapist, physiotherapist, rehabilitation worker, social worker	
Care and education	Care by:	Mother, father, parents, family, foster family, care institute, other and unknown
	Education:	Nursery school, playgroup, primary school, secondary school, other and unknown
Vaccinations date	DKTP1, DKTP2, DKTP3, DKTP4, HIB1, HIB2, HIB3, HIB4, BMR, BCG, PNCV, influenza, meningitis C, pneumovax, other	
Lab results	*HIV virology: DNA* Value (positive or negative), laboratory, sample date for the following determinates: HIV-1 DNA, HIV-2 DNA, HIV-1 antibodies, HIV-2 antibodies	
	Chemistry: The following determinates are always collected: Glucose, Amylase, ALAT/SGPT,ASAT/SGOT, Alkaline phosphatase, Gamma GT, Lactate, Triglycerides, Cholesterol, Cholesterol, HBA1c	
	Haematology: The following determinates are always collected: Haemoglobin, Leukocytes, Thrombocytes, MCV	
	Other viral infections Value (positive or negative), laboratory, sample date for the following determinates: In addition to Hepatitis and CMV, Toxoplasmosis and Varicella Zoster Virus are collected.	

Appendix K

TESSy - The HIV/AIDS metadata set

Record type	Record type version	Variable	Full name	Description	Type	Coded value list	Required	Repeatable	Allow NA	Allow UNK	Default (if not reported)	Date: Allowed formats	Number: Min value	Number: Max value
AIDS	3	RecordId	RecordId	Unique identifier for each record within and across the national surveillance system – MS selected and generated	TEXT		True (Error)	No	No	No				
AIDS	3	RecordType	RecordType	Structure and format of the data (case based reporting and aggregate reporting).	CV	'AIDS'	True (Error)	No	No	No				
AIDS	3	RecordTypeVersion	RecordTypeVersion	There may be more than one version of a recordType. This element indicates which version the sender uses when generating the message. Required when no metadata set is provided at upload.	NUM		No	No	No	No				
AIDS	3	Subject	Subject	Disease to report	CV	[Subjects for AIDS]: AIDS = AIDS	True (Error)	No	No	No				
AIDS	3	Status	Status	Status of reporting NEW/UPDATE or DELETE (inactivate). Default if left out: NEW/UPDATE. If set to DELETE, the record with the given recordId will be deleted from the TESSy database (or better stated, invalidated. If set to NEW/UPDATE or left empty, the record is newly entered into the database.	CV	[Statuses]: DELETE = Delete a previously reported record. NEW/UPDATE = Report a new or update a previously reported record (default).	No	No	No	No				
AIDS	3	DataSource	DataSource	The data source (surveillance system) that the record originates from.	CV	[Data sources] (see the coded values list)	True (Error)	No	No	No				
AIDS	3	ReportingCountry	ReportingCountry	The country reporting the record.	CV	[Countries] (see the coded values list)	True (Error)	No	No	No				
AIDS	3	DateUsedForStatistics	DateUsedForStatistics	The reference date used for standard reports that is compared to the reporting period. The date used for statistics can be any date that the reporting country finds applicable, e.g. date of notification, date of diagnosis or any other date.	DATE		True (Error)	No	No	No		yyyy, yyyy-Qq, yyyy-mm, yyyy-Www, yyyy-mm-dd		
AIDS	3	Age	Age	Exact age at diagnosis of AIDS. Age as a crude number is preferred - calculated from date of diagnosis	NUM		True (Warning)	No	No	Yes	UNK		0	120
AIDS	3	AgeClass	Age class	For AIDS case based reporting, the exact age is preferred but aggregate age classes are allowed. Previous coding as 'unknown/paediatric' and 'unknown/adult' is not allowed. The previously used age grouping is provided in the list of age groups.	CV	AgeClass2 (see the coded values list)	True (Remark)	No	No	Yes	UNK			
AIDS	3	AIDSIndicatorDisease	AIDS indicator disease at the time of AIDS diagnosis	Note that this is a repetable field and can be repeated to submit mutlipel diagnoses.	CV	AIDSIndicatorDisease: 1 = Bacterial infections, multiple or recurrent in a child under 13 years of age 10 = Histoplasmosis, disseminated or extrapulmonary 11 = Isosporiasis, intestinal with diarrhoea (>1 months duration) 12 = Mycobacterium avium complex or M. kansasii, disseminated or extrapulmonary 13 = Mycobacterium tuberculosis, pulmonary in an adult or an adolescent (aged 13 years or over)* 14 = Mycobacterium tuberculosis, extrapulmonary 15 = Mycobacterium, other species or unidentified species, disseminated or extrapulmonary 16 = Pneumocystis carinii pneumonia	True (Error)	Yes	No	No				

					17 = Pneumonia, recurrent in an adult or an adolescent (aged 13 years or over)* 18 = Progressive multifocal leukoencephalopathy 19 = Salmonella (non typhoid) septicaemia, recurrent 2 = Candidiasis of bronchi, trachea, or lungs 20 = Toxoplasmosis of brain in a patient over one month of age 21 = Cervical cancer, invasive in an adult or an adolescent (aged 13 years or over)* 22 = Encephalopathy, HIV-related 23 = Kaposi s sarcoma 24 = Lymphoid interstitial pneumonia in a child under 13 years of age 25 = Lymphoma, Burkitt s (or equivalent term) 26 = Lymphoma, immunoblastic (or equivalent term) 27 = Lymphoma, primary, of brain 28 = Wasting syndrome due to HIV 3 = Candidiasis, oesophageal 30 = Opportunistic infection(s), not specified 31 = Lymphoma(s), not specified 4 = Coccidioidomycosis, disseminated or extrapulmonary 5 = Cryptococcosis, extrapulmonary 6 = Cryptosporidiosis, intestinal with diarrhoea (>1 months duration) 7 = Cytomegalovirus disease (other than liver, spleen, or nodes) in a patient over one month of age 8 = Cytomegalovirus retinitis (with loss of vision) 9 = Herpes simplex: chronic ulcer(s) (>1 months duration); or bronchitis, pneumonitis, or oesophagitis in a patient over one month of age					
AIDS	3	ARTTreatment	Any kind of antiretroviral treatment at any time prior to AIDS	This variable indicates if the patient receives any kind of antiretroviral treatment at any time prior to AIDS diagnosis. There is no minimum duration and no restriction on timing.	CV	YesNoUnk: N=N0; Unk = Unknown; Y = Yes	True (Warning)	No	No	Yes	UNK
AIDS	3	Classification	Case Classification	For AIDS, only 'Confirmed' cases CONF are reportable at European level.	CV	CaseClassification: CONF = Confirmed POSS = Possible PROB = Probable Unk = Unknown	True (Warning)	No	No	Yes	UNK
AIDS	3	ClinicalCriteria	Clinical Criteria Met	For AIDS, only cases meeting the clinical criteria should be reported to the European level.	CV	YesNoNAUnk: N = No NA = Not applicable Unk = Unknown Y = Yes	True (Warning)	No	Yes	Yes	UNK
AIDS	3	CountryOfBirth	Country of birth of	This is the prefered variable. If it is not available, then	CV	Country_Incl_HistCountries	FALSE	No	No	Yes	UNK

			patient	submit CountryOfNationality, and if this is not available,		(see the coded values list)						
AIDS	3	CountryOfNationality	Country of nationality of patient	CountryOfBirth is the prefered variable. If it is not available, then submit CountryOfNationality, and if this is not available, then RegionOfOrigin. Please still submit CountryOfNationality, if possible.	CV	Country (see the coded values list)	FALSE	No	No	Yes		
AIDS	3	DateOfDeath	Date of death	Date of death because of HIV/AIDS	DATE		FALSE	No	Yes	Yes	UNK	yyyy, yyyy-Qq, yyyy-mm, yyyy-mm-dd, UNK, NA
AIDS	3	DateOfDiagnosis	Date of Diagnosis	Date should be provided as exact date or incomplete date. The exact date is preferred and should be provided if available; incomplete dates (e.g. quarter, month, year) are allowed as well.	DATE		True (Error)	No	No	No		yyyy, yyyy-Qq, yyyy-mm, yyyy-Www, yyyy-mm-dd
AIDS	3	DateOfHIVDiagnosis	Date of first positive HIV test	Date of first positive HIV test	DATE		FALSE	No	No	Yes	UNK	yyyy, yyyy-Qq, yyyy-mm, yyyy-mm-dd, UNK
AIDS	3	DateOfNotification	Date of Notification	The exact date is preferred and should be provided if available; incomplete dates (e.g. quarter, month, year) are allowed as well.	DATE		FALSE	No	No	Yes	UNK	yyyy, yyyy-Qq, yyyy-mm, yyyy-Www, yyyy-mm-dd, UNK
AIDS	3	DateOfOnset	Date of Onset of Disease	For AIDS, this should be coded as Unk 'Unknown' except for acute primo-infection or proven seroconversion by laboratory confirmation.	DATE		FALSE	No	No	Yes	UNK	yyyy, yyyy-Qq, yyyy-mm, yyyy-mm-dd, UNK
AIDS	3	DateOfReportDeath	Date of death report to national HIV/AIDS surveillance	The exact date is preferred to obtain more accurate information and to allow better comparison and grouping.	DATE		FALSE	No	Yes	Yes	UNK	yyyy, yyyy-Qq, yyyy-mm, yyyy-mm-dd, UNK, NA
AIDS	3	EpiLinked	Epidemiological Link Criteria Met	For AIDS, this is not applicable and should be coded as 'Not Applicable' NA.	CV	YesNoNAUnk: N = No NA = Not applicable Unk = Unknown Y = Yes	FALSE	No	Yes	Yes	UNK	
AIDS	3	Gender	Gender	Transsexual should be coded as O - Other	CV	Sex: F = Female M = Male O = Other (e.g., transsexual) Unk = Unknown	True (Warning)	No	No	Yes	UNK	
AIDS	3	HIVType	Describes the type of HIV infection	Describes the type of HIV infection	CV	HIVType: HIV1 = HIV1 only HIV12 = HIV1 and HIV2 (co-infection) HIV2 = HIV2 only Unk = Unknown	True (Error)	No	No	Yes		
AIDS	3	LaboratoryResult	Laboratory result	Following EU case definition 2008, only laboratory confirmed cases are reported. Historical data as well as cases reported using another case definition decribed in Data source may have value NA.	CV	LabResults: CONF = Confirmed NA = Not Applicable PROB = Probable UNK = Unknown	FALSE	No	Yes	Yes	UNK	
AIDS	3	Outcome	Outcome of case	Information on whether the case is alive or deceased. The death should be due to the reported disease.	CV	Outcome2: A = Alive D = Died UNK = Unknown	FALSE	No	No	Yes	UNK	
AIDS	3	PlaceOfNotification	Place of	Place of the first notification of the case to a regional	CV	NUTS	FALSE	No	No	Yes		

		notification	authority. Select the most detailed NUTS level possible.		(see the coded values list)					
AIDS	3 PlaceOfResidence	Place of residence	Place of residence of patient at the time of disease onset. Select the most detailed NUTS level possible.	CV	NUTS (see the coded values list)	FALSE	No	No	Yes	
AIDS	3 RegionOfOrigin	Region of origin of patient	Region of origin, where from patient is originating. If country of birth is reported, it should correspond to the region. CountryOfBirth is the prefered variable. If it is not available, then submit CountryOfNationality, and if this is not available, then RegionOfOrigin. Please still submit RegionOfOrigin, if possible.	CV	RegionOfOrigin: ABROAD = Abroad but sub continent unknown AUSTNZ = Australia and New Zealand CAR = Caribbean CENTEUR = Central Europe EASTASIAPAC = East Asia and Pacific EASTEUR = East Europe EUROPE = If a case can not be reported in West, central or Eastern Europe, he/she should be reported in Europe (sub-continent unknown). LATAM = Latin America NORTHAFRMIDEAST = North Africa and Middle East NORTHAM = North America REPCOUNTRY = Same as country of report SOUTHASIA = South and South East Asia SUBAFR = Sub Sahara Africa Unk = Unknown WESTEUR = West Europe	True (Warning)	No	No	Yes	UNK
AIDS	3 Transmission	Describes the most probable route of Transmission	Nosocomial infection includes patients infected in health care settings. Case of occupational exposure should be classified as UNK 'Unknown or undetermined'. Cases which are not fully documented should be coded as Unk.	CV	Transmission: HAEMO = haemophiliac patient HETERO = heterosexual contact IDU = ever injected drugs MSM = MSM/homo or bisexual male MTCT = mother-to-child-transmisison NOSO = Nosocomial TRANSFU = transfusion recipient Unk = Unknown or undetermined	True (Warning)	No	No	Yes	UNK
AIDS	3 TransmissionHetero	Specifies the heterosexual route of transmission	This allows to specify the heterosexual route of transmission and this should be provided if Transmission= HETERO (variable 21); in other cases the variable is coded as default NA 'not applicable'. Heterosexual contact refers to a person for who risk factors for HIV infection other than heterosexual contact have not been identified and who either originates from a country with generalized epidemic (HIV prevalence in pregnant women>1%, see list of countries in the Annex 3) or has had sex with: bisexual male, IDU, haemophiliac - recipient, a person from a country with generalized epidemic, a person known to be HIV positive and not known to belong to one of the above mentioned or is strongly believed to have been infected through sexual transmission, information on risk factors and HIV status of partner(s) not available.	CV	TransmissionHetero: NA = not applicable ORIGINHP = Originating from a country with generalized epidemic SEXBI = Sex with a bisexual male SEXHAEMO = Sex with a haemophiliac SEXHIVPOS = Sex with a person known to be infected and not know to belong to any of categories above SEXHP = Sex with a person originating or living in a country with a generalized epidemic SEXIDU = Sex with a injecting drug user SEXUNK = Strongly believed to have been infected through heterosexual contact, information on risk factor and partner not available	FALSE	No	Yes	No	
AIDS	3 TransmissionMTCT	TransmissionMTCT	This allows to specify the transmission categories for mother to child cases and this should be provided if Transmission=MTCT (variable 21); in other cases the variable is coded as default NA 'not applicable'.	CV	TransmissionMTCT: MOTHHET = Infected through heterosexual contact and not known to belong to category above	FALSE	No	Yes	Yes	

					MOTHHP = Originating from a country with generalized epidemic MOTHIDU = Injecting drug use MOTHTRANSFU = Transfusion recipient NA = not applicable Unk = Other/undetermined								
HIV	4	RecordId	RecordId	Unique identifier for each record within and across the national surveillance system – MS selected and generated	TEXT		True (Error)	No	No	No			
HIV	4	RecordType	RecordType	Structure and format of the data (case based reporting and aggregate reporting).	CV	'HIV'	True (Error)	No	No	No			
HIV	4	RecordTypeVersion	RecordTypeVersion	There may be more than one version of a recordType. This element indicates which version the sender uses when generating the message. Required when no metadata set is provided at upload.	NUM		No	No	No	No			
HIV	4	Subject	Subject	Disease to report	CV	[Subjects for HIV]: HIV = HIV infection	True (Error)	No	No	No			
HIV	4	Status	Status	Status of reporting NEW/UPDATE or DELETE (inactivate). Default if left out: NEW/UPDATE. If set to DELETE, the record with the given recordId will be deleted from the TESSy database (or better stated, invalidated. If set to NEW/UPDATE or left empty, the	CV	[Statuses]: DELETE = Delete a previously reported record. NEW/UPDATE = Report a new or update a previously reported record (default).	No	No	No	No			
HIV	4	DataSource	DataSource	The data source (surveillance system) that the record originates from.	CV	[Data sources] (see the coded values list)	True (Error)	No	No	No			
HIV	4	ReportingCountry	ReportingCountry	The country reporting the record.	CV	[Countries] (see the coded values list)	True (Error)	No	No	No			
HIV	4	DateUsedForStatistics	DateUsedForStatistics	The reference date used for standard reports that is compared to the reporting period. The date used for statistics can be any date that the reporting country finds applicable, e.g. date of notification, date of diagnosis or any other date.	DATE		True (Error)	No	No	No		yyyy, yyyy-Qq, yyyy-mm, yyyy-Www, yyyy-mm-dd	
HIV	4	Age	Age	Exact age at diagnosis of HIV. Age as a crude number is preferred - calculated from date of diagnosis	NUM		True (Warning)	No	No	Yes	UNK	0	120
HIV	4	CD4Cells	CD4 count at the time of HIV diagnosis	CD4 count at the time of HIV diagnosis	NUM		FALSE	No	Yes	Yes		0	6000
HIV	4	Classification	Case Classification	For HIV, only 'Confirmed' cases CONF are reportable at European level. In case when MTCT cases <18 month are not confirmed, they should be reported as probable "PROB".	CV	CaseClassification: CONF = Confirmed POSS = Possible PROB = Probable Unk = Unknown	True (Warning)	No	No	Yes	UNK		
HIV	4	ClinicalCriteria	Clinical Criteria Met	For HIV, NA is the expected value.	CV	YesNoNAUnk: N = No NA = Not applicable Unk = Unknown Y = Yes	FALSE	No	Yes	Yes	UNK		
HIV	4	CountryOfBirth	Country of birth of patient	This is the prefered variable. If it is not available, then submit CountryOfNationality, and if this is not available,	CV	Country_Incl_HistCountries (see the coded values list)	FALSE	No	No	Yes			
HIV	4	CountryOfNationality	Country of nationality of	CountryOfBirth is the prefered variable. If it is not available, then submit CountryOfNationality, and if this	CV	Country (see the coded values list)	FALSE	No	No	Yes			
HIV	4	DateOfAIDSDiagnosis	Date of AIDS diagnosis	For HIV cases initially reported at a pre-AIDS stage, the date of AIDS diagnosis is 'follow-up' information, which necessitates updating of the record. The exact date is preferred to obtain more accurate information and to allow better comparison and grouping. Incomplete dates (quarter, month, year) are allowed as well.	DATE		FALSE	No	Yes	Yes		yyyy, yyyy-Qq, yyyy-mm, yyyy-mm-dd, UNK, NA	

HIV	4	DateOfDeath	Date of death	Date of death because of HIV/AIDS	DATE		FALSE	No	Yes	Yes	UNK	yyyy, yyyy-Qq, yyyy-mm, yyyy-mm-dd, UNK, NA
HIV	4	DateOfDiagnosis	Date of Diagnosis	Date should be provided as exact date or incomplete date. The exact date is preferred and should be provided if available; incomplete dates (e.g. quarter, month, year) are allowed as well.	DATE		True (Error)	No	No	No		yyyy, yyyy-Qq, yyyy-mm, yyyy-Www, yyyy-mm-dd
HIV	4	DateOfNotification	Date of Notification	Date when the case report is first reported or notified to public health authorities.	DATE		FALSE	No	No	Yes	UNK	yyyy, yyyy-Qq, yyyy-mm, yyyy-Www, yyyy-mm-dd, UNK
HIV	4	DateOfOnset	Date of Onset of Disease	Date of onset of disease. Not applicable in asymptomatic cases. If not applicable, please use 'Unk'	DATE		FALSE	No	No	Yes	UNK	yyyy, yyyy-Qq, yyyy-mm, yyyy-mm-dd, UNK
HIV	4	EpiLinked	Epidemiological Link Criteria Met	For HIV, this shoudl be coded as NA.	CV	YesNoNAUnk: N = No NA = Not applicable Unk = Unknown Y = Yes	FALSE	No	Yes	Yes	UNK	
HIV	4	Gender	Gender	Transsexual should be coded as O - Other	CV	Sex: F = Female M = Male O = Other (e.g., transsexual) Unk = Unknown	True (Warning)	No	No	Yes	UNK	
HIV	4	HIVStatus	HIV status; previous positive test	This variable provides information on previous positive test results, prior to the first time of reporting. This variable allows cases "newly diagnosed" to be distinguished from case who had positive HIV test in the past but are reported for the first time in the country.	CV	HIVStatus: NEG = Negative PREVPOS = previous HIV positive Unk = Unknown	FALSE	No	No	Yes	UNK	
HIV	4	HIVType	Describes the type of HIV infection	Describes the type of HIV infection	CV	HIVType: HIV1 = HIV1 only HIV12 = HIV1 and HIV2 (co-infection) HIV2 = HIV2 only Unk = Unknown	True (Error)	No	No	Yes		
HIV	4	LaboratoryResult	Laboratory result	For HIV, CONF is the expected coding.	CV	LabResults: CONF = Confirmed NA = Not Applicable PROB = Probable UNK = Unknown	FALSE	No	Yes	Yes	UNK	
HIV	4	Outcome	Outcome of case	Information on whether the case is alive or deceased. The death should be due to the reported disease.	CV	Outcome2: A = Alive D = Died UNK = Unknown	FALSE	No	No	Yes	UNK	
HIV	4	PlaceOfNotification	Place of notification	Place of the first notification of the case to a regional authority. Select the most detailed NUTS level possible.	CV	NUTS (see the coded values list)	FALSE	No	No	Yes		
HIV	4	PlaceOfResidence	Place of residence	Place of residence of patient at the time of disease onset. Select the most detailed NUTS level possible.	CV	NUTS (see the coded values list)	FALSE	No	No	Yes		
HIV	4	ProbableCountryOfInfection	Probable Country Of Infection	If Imported=Yes: One entry for each country visited during the incubation period of the disease should be provided. The entry could be UNK even if the case is known to be imported.	CV	Country_Incl_HistCountries (see the coded values list)	FALSE	No	No	Yes		

HIV	4	RegionOfOrigin	Region of origin of patient	Region of origin, where from patient is originating. If country of birth is reported, it should correspond to the region. CountryOfBirth is the prefered variable. If it is not available, then submit CountryOfNationality, and if this is not available, then RegionOfOrigin. Please still submit RegionOfOrigin, if possible.	CV	RegionOfOrigin: ABROAD = Abroad but sub continent unknown AUSTNZ = Australia and New Zealand CAR = Caribbean CENTEUR = Central Europe EASTASIAPAC = East Asia and Pacific EASTEUR = East Europe EUROPE = If a case can not be reported in West, central or Eastern Europe, he/she should be reported in Europe (sub-continent unknown). LATAM = Latin America NORTHAFRMIDEAST = North Africa and Middle East NORTHAM = North America REPCOUNTRY = Same as country of report SOUTHASIA = South and South East Asia SUBAFR = Sub Sahara Africa Unk = Unknown WESTEUR = West Europe	True (Warning)	No	No	Yes	UNK
HIV	4	Stage	Clinical stage at the time of HIV diagnosis	This variable specifies the clinical stage at the time of HIV diagnosis. In case of acute infection (Stage=ACUTE) the DateOfOnset should be provided. • In countries with both laboratory and clinician reports, the latter may be delayed and the clinical stage may evolve in the meantime. In such cases, the clinical stage should be that one provided by the clinician.	CV	Stage: ACUTE = Acute HIV infection AIDS = AIDS ASYMP = Asymptomatic NONAIDS = Non-AIDS, not further specified SYMPNONAIDS = Symptomatic non-AIDS UNK = Unknown	FALSE	No	No	Yes	UNK
HIV	4	Transmission	Describes the most probable route of Transmission	Nosocomial infection includes patients infected in health care settings. Case of occupational exposure should be classified as UNK 'Unknown or undetermined'. Cases which are not fully documented should be coded as UNK.	CV	Transmission: HAEMO = haemophiliac patient HETERO = heterosexual contact IDU = ever injected drugs MSM = MSM/homo or bisexual male MTCT = mother-to-child-transmisison NOSO = Nosocomial TRANSFU = transfusion recipient Unk = Unknown or undetermined	True (Warning)	No	No	Yes	UNK
HIV	4	TransmissionHetero	Specifies the heterosexual route of transmission	This allows to specify the heterosexual route of transmission and this should be provided if Transmission=HETERO (variable 21); in other cases the variable is coded as default NA 'not applicable'. Heterosexual contact refers to a person for who risk factors for HIV infection other than heterosexual contact have not been identified and who either originates from a country with generalized epidemic (HIV prevalence in pregnant women>1%) or has had sex with: bisexual male, IDU, haemophiliac - recipient, a person from a country with generalized epidemic, a person known to be HIV positive and not known to belong to one of the above mentioned or is strongly believed to have been infected through sexual transmission, information on risk factors and HIV status of partner(s) not available. The sub-variable heterosexual transmission includes not only information on the reported.	CV	TransmissionHetero: NA = not applicable ORIGINHP = Originating from a country with generalized epidemic SEXBI = Sex with a bisexual male SEXHAEMO = Sex with a haemophiliac SEXHIVPOS = Sex with a person known to be infected and not know to belong to any of categories above SEXHP = Sex with a person originating or living in a country with a generalized epidemic SEXIDU = Sex with a injecting drug user SEXUNK = Strongly believed to have been infected through heterosexual contact, information on risk factor and partner not available	FALSE	No	Yes	No	

HIV	4	TransmissionMTCT	TransmissionMTCT	This allows to specify the transmission categories for mother to child cases and this should be provided if Transmission=MTCT (variable 21); in other cases the variable is coded as default NA 'not applicable'.	CV	TransmissionMTCT: MOTHHET = Infected through heterosexual contact and not known to belong to category above MOTHHP = Originating from a country with generalized epidemic MOTHIDU = Injecting drug use MOTHTRANSFU = Transfusion recipient NA = not applicable Unk = Other/undetermined	FALSE	No	Yes	Yes		
HIVAIDS	1	RecordId	RecordId	Unique identifier for each record within and across the national surveillance system – MS selected and generated	TEXT		True (Error)	No	No	No		
HIVAIDS	1	RecordType	RecordType	Structure and format of the data (case based reporting and aggregate reporting).	CV	'HIVAIDS'	True (Error)	No	No	No		
HIVAIDS	1	RecordTypeVersion	RecordTypeVersion	There may be more than one version of a recordType. This element indicates which version the sender uses when generating the message. Required when no metadata set is provided at upload.	NUM		No	No	No	No		
HIVAIDS	1	Subject	Subject	Disease to report	CV	[Subjects for HIVAIDS]: HIVAIDS = HIV diagnoses case-based, including AIDS	True (Error)	No	No	No		
HIVAIDS	1	Status	Status	Status of reporting NEW/UPDATE or DELETE (inactivate). Default if left out: NEW/UPDATE. If set to DELETE, the record with the given recordId will be deleted from the TESSy database (or better stated, invalidated. If set to NEW/UPDATE or left empty, the record is newly entered into the database.	CV	[Statuses]: DELETE = Delete a previously reported record. NEW/UPDATE = Report a new or update a previously reported record (default).	No	No	No	No		
HIVAIDS	1	DataSource	DataSource	The data source (surveillance system) that the record originates from.	CV	[Data sources] (see the coded values list)	True (Error)	No	No	No		
HIVAIDS	1	ReportingCountry	ReportingCountry	The country reporting the record.	CV	[Countries] (see the coded values list)	True (Error)	No	No	No		
HIVAIDS	1	DateUsedForStatistics	DateUsedForStatistics	The reference date used for standard reports that is compared to the reporting period. The date used for statistics can be any date that the reporting country finds applicable, e.g. date of notification, date of diagnosis or any other date.	DATE		True (Error)	No	No	No	yyyy, yyyy-Qq, yyyy-mm, yyyy-Www, yyyy-mm-dd	
HIVAIDS	1	AcuteInfection	Evidence of recent infection, aside from the recent infection assay result	An infection can be considered to be recent if a patient demonstrates with seroconversion illness, has a negative HIV test within 6 months of diagnosis or has evidence from p24 antigen or Western Blot assays. This is a repeatable field and up to 4 options can be entered. You should replace empty fields with N/A.	CV	AcuteInfectionHIV: EV24ANT = Evidence from p24 antigen EVWBLOT = Evidence from Western Blot NA = Not applicable (not acute infection) NEGTEST = Last negative test within 6 months of HIV diagnosis SEROILL = Seroconversion illness UNK = Unknown	FALSE	Yes	Yes	Yes		
HIVAIDS	1	Age	Age	Exact age at diagnosis of HIV. Age as a crude number is preferred. If not available, age can be calculated by the MS from the date of birth and date of diagnosis.	NUM		True (Warning)	No	No	Yes	0	120
HIVAIDS	1	AIDSIndicatorDisease	AIDS indicator disease at the time of AIDS diagnosis	AIDS indicator disease at the time of AIDS diagnosis occurring within two consecutive months from the date of AIDS diagnosis. This is a repeatable field for up to 4 diagnoses. You should replace empty fields with N/A.	CV	AIDSIndicatorDisease: 1 = Bacterial infections, multiple or recurrent in a child under 13 years of age 10 = Histoplasmosis, disseminated or extrapulmonary	FALSE	Yes	No	No		

				11 = Isosporiasis, intestinal with diarrhoea (>1 months duration) 12 = Mycobacterium avium complex or M. kansasii, disseminated or extrapulmonary 13 = Mycobacterium tuberculosis, pulmonary in an adult or an adolescent (aged 13 years or over)* 14 = Mycobacterium tuberculosis, extrapulmonary 15 = Mycobacterium, other species or unidentified species, disseminated or extrapulmonary 16 = Pneumocystis carinii pneumonia 17 = Pneumonia, recurrent in an adult or an adolescent (aged 13 years or over)* 18 = Progressive multifocal leukoencephalopathy 19 = Salmonella (non typhoid) septicaemia, recurrent 2 = Candidiasis of bronchi, trachea, or lungs 20 = Toxoplasmosis of brain in a patient over one month of age 21 = Cervical cancer, invasive in an adult or an adolescent (aged 13 years or over)* 22 = Encephalopathy, HIV-related 23 = Kaposi s sarcoma 24 = Lymphoid interstitial pneumonia in a child under 13 years of age 25 = Lymphoma, Burkitt s (or equivalent term) 26 = Lymphoma, immunoblastic (or equivalent term) 27 = Lymphoma, primary, of brain 28 = Wasting syndrome due to HIV 3 = Candidiasis, oesophageal 30 = Opportunistic infection(s), not specified 31 = Lymphoma(s), not specified 4 = Coccidioidomycosis, disseminated or extrapulmonary 5 = Cryptococcosis, extrapulmonary 6 = Cryptosporidiosis, intestinal with diarrhoea (>1 months duration) 7 = Cytomegalovirus disease (other than liver, spleen, or nodes) in a patient over one month of age 8 = Cytomegalovirus retinitis (with loss of vision) 9 = Herpes simplex: chronic ulcer(s) (>1 months duration); or bronchitis, pneumonitis, or oesophagitis in a patient over one month of age				
HIVAIDS	1 ART	Antiretroviral therapy	Was the patient receiving antiretroviral therapy (ART) at the date last seen for care? If yes, last attendance date should be reported.	CV	YesNoUnk: N = No Unk = Unknown Y = Yes	FALSE	No	No	Yes

HIVAIDS	1	CD4Latest	Last known CD4 count	Last known CD4 count. This variable is for historical updates on CD4 provided over time. Enter the numeric value of the CD4 (0-6000) or unknown (UNK).	NUM		FALSE	No	No	Yes			0	6000
HIVAIDS	1	CD4LatestDate	Date of last CD4 count assessment	The exact date is preferred and should be provided if available; incomplete dates (e.g. week, quarter, month, year) are allowed if exact date is not available. If CD4Latest is not available, enter NA for date.	DATE		FALSE	No	Yes	Yes		yyyy, yyyy-Qq, yyyy-mm, yyyy-Www, yyyy-mm-dd, UNK, NA		
HIVAIDS	1	CountryOfBirth	Country of birth of patient	Defines the country of birth at country level, the ISO list of countries is provided. CountryofBirth is the preferred variable. If Unknown code as UNK.	CV	Country_Incl_HistCountries (see the coded values list)	True (Warning)	No	No	Yes				
HIVAIDS	1	DateOfAIDSDiagnosis	Date of AIDS diagnosis	Date of AIDS diagnosis.The exact date is preferred and should be provided if available; incomplete dates (e.g. week, quarter, month, year) are allowed if exact date is not available.	DATE		FALSE	No	Yes	Yes		yyyy, yyyy-Qq, yyyy-mm, yyyy-Www, yyyy-mm-dd,		
HIVAIDS	1	DateOfDeath	Date of death	Date of death due to any cause. The exact date is preferred to obtain more accurate information and to allow better comparison and grouping. Incomplete dates (quarter, month, year) are permissible. All cases that are still alive or where the outcome (i.e., whether the case is alive or dead) is unknown are to be coded as 'NA'.	DATE		True (Warning)	No	Yes	Yes	NA	yyyy, yyyy-Qq, yyyy-mm, yyyy-Www, yyyy-mm-dd, UNK, NA		
HIVAIDS	1	DateOfDiagnosis	Date of Diagnosis	The date of first HIV diagnosis; clinical or laboratory diagnosis. Date should be provided as exact date or incomplete date. The exact date is preferred and should be provided if available; incomplete dates (e.g. quarter, month, year) are allowed if exact date is not available.	DATE		True (Error)	No	No	No		yyyy, yyyy-Qq, yyyy-mm, yyyy-Www, yyyy-mm-dd		
HIVAIDS	1	DateOfNotification	Date of Notification	This is the date on which the HIV case was notified for the first time to the reporting country.The exact date is preferred and should be provided if available; incomplete dates (e.g. quarter, month, year) are allowed if exact date is not available.	DATE		True (Warning)	No	No	Yes		yyyy, yyyy-Qq, yyyy-mm, yyyy-Www, yyyy-mm-dd, UNK		
HIVAIDS	1	DeathCause	Outcome of case	Information on whether the case is alive or deceased (due to AIDS-related and non-AIDS related causes).	CV	DeathCause: A = Alive DAIDS = Death due to AIDS DNOAIDS = Non AIDS-related death DUNK = Died of unknown cause UNK = Unknown	True (Warning)	No	No	Yes				
HIVAIDS	1	FirstCD4Count	CD4 cell count at time of diagnosis	The variable specifies the CD4 cells count at the time of HIV diagnosis. Enter the numeric value of the CD4 (0-6000) or unknown (UNK).	NUM		True (Warning)	No	No	Yes			0	6000
HIVAIDS	1	FirstCD4Date	Date of first CD4 cell count at time of diagnosis	The exact date is preferred and should be provided if available; incomplete dates (e.g. week, quarter, month, year) are allowed if exact date is not available. If FirstCD4Count is not available, enter NA for date.	DATE		True (Warning)	No	Yes	Yes		yyyy, yyyy-Qq, yyyy-mm, yyyy-Www, yyyy-mm-dd, UNK, NA		
HIVAIDS	1	Gender	Gender	Transsexual should be coded as O - Other	CV	Sex: F = Female M = Male O = Other (e.g., transsexual) Unk = Unknown	True (Warning)	No	No	Yes				
HIVAIDS	1	HIVStatus	HIV status; previous positive	This variable provides information on previous positive test results, prior to the first time of reporting. This	CV	HIVStatus: NEG = Negative	FALSE	No	No	Yes				

		test	variable allows cases "newly diagnosed" to be distinguished from case who had positive HIV test in the past but are reported for the first time in the country.		PREVPOS = previous HIV positive Unk = Unknown					
HIVAIDS	1	HIVType	Describes the type of HIV infection	Describes the type of HIV infection	CV	HIVType: HIV1 = HIV1 only HIV12 = HIV1 and HIV2 (co-infection) HIV2 = HIV2 only Unk = Unknown	True (Error)	No	No	Yes	
HIVAIDS	1	LastAttendanceDate	Date the patient was last seen for HIV-related care (can be a date prior to the	The exact date is preferred and should be provided if available; incomplete dates (e.g. week, quarter, month, year) are allowed if exact date is not available.	DATE		FALSE	No	No	Yes	yyyy, yyyy-Qq, yyyy-mm, yyyy-Www, yyyy-mm-dd, UNK
HIVAIDS	1	ProbableCountryOfInfection	Probable country of infection	Country or countries where infection of the patient is likely to have occurred.	CV	Country_Incl_HistCountries (see the coded values list)	FALSE	Yes	No	Yes	
HIVAIDS	1	RegionOfOrigin	Region of origin of patient	Region from which the case originates. If the case is from the reporting country, it should be coded as REPCOUNTRY. CountryOfBirth is the prefered variable. If this is not available, then submit RegionOfOrigin. If both are submitted, CountryOfBirth should match RegionOfOrigin.	CV	RegionOfOrigin: ABROAD = Abroad but sub continent unknown AUSTNZ = Australia and New Zealand CAR = Caribbean CENTEUR = Central Europe EASTASIAPAC = East Asia and Pacific EASTEUR = East Europe EUROPE = If a case can not be reported in West, central or Eastern Europe, he/she should be reported in Europe (sub-continent unknown). LATAM = Latin America NORTHAFRMIDEAST = North Africa and Middle East NORTHAM = North America REPCOUNTRY = Same as country of report SOUTHASIA = South and South East Asia SUBAFR = Sub Sahara Africa Unk = Unknown WESTEUR = West Europe	FALSE	No	No	Yes	
HIVAIDS	1	Transmission	Describes the most probable route of HIV transmission classified as:	Most probable route of HIV transmission classified as: sex between men; heterosexual contact; ever injected drugs; mother-to-child transmission; transfusion recipient; nosocomial. Nosocomial infection includes patients infected in health care settings. Case of occupational exposure should be classified as exposure unknown or undetermined. Cases which are not fully documented should also be coded as unknown.	CV	Transmission: HAEMO = haemophiliac patient HETERO = heterosexual contact IDU = ever injected drugs MSM = MSM/homo or bisexual male MTCT = mother-to-child-transmisison NOSO = Nosocomial TRANSFU = transfusion recipient Unk = Unknown or undetermined	True (Warning)	No	No	Yes	
HIVAIDS	1	TransmissionPartner	Describes the most probable route of HIV transmission of the partner	Describes the most probable route of HIV transmission of the partner. A list of countries with generalised HIV epidemics can be found in the reporting protocol.	CV	TransmissionPartnerHIV: PHAEMO = Partner haemophiliac PHETEPI = Partner heterosexual from generalised epidemic country PHETNEPI = Partner heterosexual from non-generalised epidemic country PIBLOOD = Partner infected through blood products PIDU = Partner injecting drug user PINOSO = Partner infected nosocomially	FALSE	No	No	Yes	

					PIVER = Partner infected through mother-to-child transmission PMSM = Partner MSM UNK = Partner undetermined or unknown					
HIVAIDS	1 VLLatest	Last known viral load	Last known viral load. Enter the numeric value. If the test provides no numeric value and is simply "low" or "undetectable" code as 0.If the VL is greater than 999999 mm3, code as 999999. If the latest viral load is unknown code as UNK.	NUM		FALSE	No	No	Yes	
HIVAIDS	1 VLLatestDate	Date of last known viral load assessment (date of blood test where available)	The exact date is preferred and should be provided if available; incomplete dates (e.g. week, quarter, month, year) are allowed if exact date is not available. If VLLatest is not available, enter NA for date.	DATE		FALSE	No	Yes	Yes	yyyy, yyyy-Qq, yyyy-mm, yyyy-Www, yyyy-mm-dd, UNK, NA
HIVAIDS	1 YearOfArrival	Year patient arrived in the reporting country	Year patient arrived in the reporting country.	DATE		FALSE	No	Yes	Yes	yyyy, UNK, NA